SAPIENZA
UNIVERSITÀ DI ROMA

Sapienza University of Rome

Department of Statistical Sciences
Ph.D. in School of Statistical Sciences

This work is submitted in partial fulfillment of the requirement for the Ph.D.

Prediction of acute kidney injury using the Electronic Medical Records of a pediatric cardiac intensive care unit

Thesis Advisors
Prof. Luca Tardella

Candidate
Davide Passaro
matricola 685572

Academic Year MMXIX-MMXXII (XXXV cycle)

A mio figlio Mattia,

A tutti i bambini incontrati in terapia intensiva

.

Abstract

Acute Kidney Injury (AKI) is a frequent complication in hospitalized patients significantly associated with mortality, length of stay, and healthcare cost. Management of AKI presents an important challenge and clinicians may be helped by robust prediction models for risk evaluation, foster prevention, and recognition. The advances in clinical informatics and the increasing availability of electronic medical records (EMR) have favored the development of predictive models of risk estimation in AKI.

In this dissertation, we analyze the problem of predicting the AKI stage during the patient's stay in the intensive care unit using retrospectively the Electronic medical records (EMRs) recently introduced in the Pediatric Intensive Care Unit (PCICU) of "Ospedale Pediatrico Bambino Gesù".

After the initial phase of data selection, extraction, and management of missing data, we develop a random forest (RF) classification model including a variable selection step with the aim of predicting the stage of AKI 48 hours in advance in both binary and multiclass cases.

The performances obtained in terms of Area under the ROC Curve (AUC-ROC) for binary cases and accuracy for multiclass cases are always very good compared with other recent attempts in the literature. The list of the most important variables obtained in the various classifications highlights the importance of some of the expected variables (such as creatinine) reported in other studies in the literature but also the presence of variables that are specific to pediatric patients under examination (such as PIM3).

Moreover, we develop other classifications using the Generalized Additive Models (GAMS) and Bayesian network (BN) models that have the benefit of offering a more interpretable approach. Although these results are inferior to the RF, they are comparable with many outcomes reported in the literature. The plot obtained with GAMs and the structure of directed acyclic graph (DAG) achieved with BN are consistent with a possible medical explanation and would present further interpretation hints for the doctors about the onset of AKI. Finally, we observe that all implemented models confirm the possibility of making an accurate prediction of the AKI stage using the PCICU. These models can be potentially included in a web interface and, in perspective, be integrated into the EMR of PCICU. This tool would allow the doctors to predict prospectively the patient's stage of AKI and evaluate how to intervene if necessary.

In order to proceed with this, it would be necessary in the future to implement the export of a larger dataset adding new data acquired in the meantime in PCICU.

Motivation

I'm a high school mathematics and physics teacher with a background as an IT consultant for a software house that operates in the field of Remote Sensing.

The motivation for joining a Ph.D. program and developing a thesis project on PCICU EMR for AKI prediction comes from the illness of my second son.

At the time of my son's second heart operation, the Pediatric Cardiac Intensive Care Unit (PCICU) of the "Bambino Gesù" hospital introduced the electronic medical record (EMR), and I saw the opportunity to do something useful not only for him but also for other children hospitalized in the intensive care unit.

I discussed the idea with the doctors and I made my skills available as a volunteer on the existing accumulated data, starting from a specific problem identified by a group of doctors in the facility.

In order to proceed with this project in the best possible way I embarked on a Ph.D. in methodological statistics at the Sapienza University of Rome. The Ph.D. program offered the opportunity to improve my methodological skills to join this challenging project of extracting useful information from the massive availability of data.

In the coming years, all hospitals will be equipped with electronic medical records and a large amount of data will be accumulated over time.

During the program, I reinforced my belief that the increasingly widespread introduction of medical records will provide an important new source of data to support physician decisions.

I realize that an important role will be played in this process by the interaction between doctors and other specialists (statisticians, data scientists, engineers, and computer scientists).

Unfortunately, having lived as a parent with the experience of intensive care for several days, I am aware of the complexity and variety of the data acquired in intensive care.

Nevertheless, the publications present in the literature over the last 20 years have shown the potential of using EMR data to support the decision of doctors and suggested that my view and convictions could be increasingly confirmed. I can not be sure that there will be a "paradigm shift" in the near future in clinical practice thanks to the use of EMR data, but in my opinion, it will be important to develop decision support tools for doctors.

In order to do this, it is important to support and incentivize the collaboration between doctors and scientists, who are capable of helping them to develop these tools.

Currently, the EMR is certainly a valid support for doctors in patient care. However, only a minimal part of the data recorded in the database is subsequently used. I hope that my work will encourage other researchers in the future to further advance this field and that the interaction between different realities (hospitals, universities, and companies that provide IT services) will lead to effective solutions capable of supporting more and more doctors in patient care.

Contents

List of Figures

List of Tables

Acronyms

AKI Acute Kidney Injury

AUC Area under the curve

BN Bayesian Network

CDSS clinical decision support systems

DAG directed acyclic graph

EHR Electronic Health Record

EMR Electronic Medical Record

GAMs Generalized Additive Models

ICU Intensive Care Unit

MAR Missing at Random

MMPC Max-Min Parents and Children algorithm

ODEs One-Dependence Estimators

OPBG Ospedale Pediatrico Bambino Gesù

PCICU Pediatric Cardiac Intensive Care Unit

RFE Recursive Feature Elimination

RWD Real World Data

SCr Serum Creatinine

TAN tree-augmented network

Introduction

This dissertation analyzes the possibility of the use of electronic medical records acquired in the PCICU of "Ospedale Pediatrico Bambino Gesù" recently introduced in this facility. In particular, following the literature and the doctors' instructions, we focused on the problem of Acute Kidney Injury (AKI) a frequent complication in hospitalized patients. We consider the problem of predicting the AKI stage during the patient's stay in the intensive care unit.

Given the type of problem and the data used, the aim of this work is mainly applicative. The adopted methodological approach is focused more on predicting the phenomenon.

The contribution of this work is essentially that of having developed a system capable of predicting the onset of AKI with a temporal advance that allows doctors to evaluate any actions, starting from raw and never-before-extracted data with this level of detail.

Below we outline in a synthetic way the contents of the different chapters that make up this thesis dissertation.

Chapter 1 presents the acute kidney injury problem, pointing out the high importance of developing methods to forecast when patients are at risk for AKI in order to improve patient outcomes.

This chapter also explains the introduction of electronic medical records (EMRs). It illustrates how the data stored in the EMR are actually used with the aim of supporting doctors in many challenging open problems in the medical field.

Finally, Chapter 1 is closed with a review of the recent attempts to predict AKI using EMR.

Chapter 2 focuses on the EMR data from the PCICU of "Ospedale Pediatrico Bambino Gesù" describing this kind of data. In particular, it describes the phase of data selection, extraction, preparation, cleansing, and management of missing data. This dissertation has a primarily applicative aim and used a methodological approach focused on prediction. The contribution of this work is essentially to have developed, starting from raw and never-before-extracted data with this level of detail, a system capable of predicting the onset of AKI with a temporal advance that allows doctors to evaluate any actions.

The specificity of this dataset has required us to make choices that, although consistent with the literature, have specific characteristics of uniqueness in particular for the type of patients involved and the selected variables.

Chapter 3 will focus on AKI forecast. It presents a random forest classification model, including a variable selection step, for the modeling of AKI in binary and multiclass cases. We also analyzed the use of an RFE algorithm in order to reduce the number of variables involved in the model.

Based on these results we chose to explore the possibility of using models that benefit from being more interpretable and offering more information about the role of the involved variable. In a sector such as the medical one, interpretability is not a negligible aspect since it can provide

possible confirmations on some clinical dynamics or further unexpected hypotheses. For this reason, in Chapters 4 and 5, we describe two alternative models applied to the problem:

- General Additive Models;
- Bayesian Networks.

As for RF, in the case of BN, we use a variable selection algorithm in order to possibly provide comparable performances with fewer covariates.

In Chapter 6 we analyze an interpretable machine learning approach in order to provide clinicians:

- a graphical representation using the partial dependence plot of each variable involved in the random forest classification models;
- a classification tree surrogate model.

Finally, in the Conclusion, we highlight the aspects that, according to our point of view, are characterizing and worthy of interest in our work.

Chapter 1

Acute Kidney Injury and Electronic Medical Records

This chapter, after a brief introduction to the problem of Acute Kidney Injury (AKI) and the use of Electronic Medical Record (EMR), analyzes the literature on the use of EMR, in order to provide clinicians with useful tools to predict the onset of AKI.

In the first paragraph of this chapter, it will be introduced the AKI definition, in the second a brief overview of Electronic Medical Records (EMRs) and in the third paragraph a review of the literature where different methods are used to predict the patient's AKI using this type of data.

1.1 Acute Kidney Injury

Acute kidney injury (AKI) is a frequent complication in hospitalized patients. The term AKI has replaced acute renal failure because smaller changes in kidney function without overt failure can result in significant clinical consequences and increased morbidity and mortality. AKI is an increasingly common clinical problem faced by nephrologists, intensivists, general physicians, and surgeons. It is also associated with adverse outcomes both in the short and long term with chronic kidney disease. AKI is significantly associated with mortality, length of stay, and healthcare cost (Chertow et al. [2005]).

In light of the impact of AKI on short and long-term outcomes, it is of high importance to develop methods to identify when patients are at risk for AKI and to diagnose subclinical AKI in order to improve patient outcomes.

Since 2012, a novel definition of AKI was proposed by the KDIGO (Kidney Disease Improving Global Outcomes) workgroup (Khwaja [2012]) in the attempt of reconciling previous formulations and adapting them to pediatric patients. This classification aims to assist practitioners caring for adults and children at risk for or with AKI, including contrast-induced acute kidney injury (CI-AKI) (Khwaja [2012]).

AKI is defined with the occurrence of at least one of the conditions reported in the following list:

(a) Increase in Serum Creatinine (SCr) by $\geq 0.3 mg/dl$ ($\geq 26.5 \mu mol/l$) within last 48 hours; or

(b) Increase in SCr to ≥ 1.5 times baseline, which is known or presumed to have occurred within

the prior 7 days; or

(c) Urine volume $< 0.5\,ml/kg/h$ for 6 hours.

AKI is also staged for severity according to criteria reported in Table 1.1.

Staging of AKI		
Stage	**Serum creatinine (SCr)**	**Urine output**
1	1.5–1.9 times baseline OR $\geq$ 0.3 mg/dl ($\geq$ 26.5 μ mol/l) increase	$<$ 0.5 ml/kg/h for 6–12 hours.
2	2.0–2.9 times baseline	$<$ 0.5 ml/kg/h for $\geq$ 12 hours.
3	3.0 times baseline OR Increase in serum creatinine to $\geq$ 4.0 mg/dl ($\geq$ 353.6 μ mol/l) OR Initiation of renal replacement therapy OR, In patients $<$ 18 years, decrease in eGFR to $<$ 35 ml/min per 1.73 m^2	$<$ 0.3 ml/kg/h for $\geq$ 24 hours OR Anuria for $\geq$ 12 hours.

Table 1.1: Staging of AKI.

This new grading is considered an advancement compared to previous criteria. In fact, this definition joint with the availability of electronic records has enabled timely alerts to warn the clinician of the occurrence and severity of AKI. Despite that there are still some open issues on this definition as highlighted in Thomas et al. [2014].

Moreover, clinicians have increasingly recognized the scale and effect of AKI, but significant challenges remain in reducing AKI incidence and improving outcomes.

Indeed AKI is a heterogeneous cluster of pathophysiologic processes that affect in the rise of serum creatinine and/or drop in urine output (UO). The diagnosis of AKI is often delayed by the reliance upon these bio-markers of kidney function that are often late and inaccurate (Soranno et al. [2022], Kaddourah et al. [2017]).

Changes in kidney function are detected by a change in biomarkers, the most common biomarker being serum creatinine (SCr). Serum creatinine is an imperfect biomarker for recognizing AKI, given that an increase in SCr often lags (48–72 hours) behind the onset of the injury. In addition, SCr is not in a steady-state condition in critically ill patients, leading to inaccurate estimates of glomerular filtration rates (eGFRs). Using an imperfect biomarker for AKI definition, recognition, and management may affect patient outcomes.

Advances in clinical informatics and the increasing availability of electronic medical records allowed the development of predictive models for risk estimation for AKI and provide a better method to predict AKI overcoming this problem of lag observed behind renal injury.

1.2 Electronic Medical Records

In the last thirty years, the development of technologies has favored the development of Electronic Medical Records (EMRs) and a more general version denoted as Electronic Health Record (EHR).

Electronic Medical Records (EMRs) are the digital version of a patient's paper chart. EMRs are real–time, patient–centered records that make information available instantly and securely to authorized users.

As reported in Mcmullen et al. [2014], EMRs and EHRs have become essential systems by which it is possible to communicate vital patient information to other members of the health care team

as well as to patients.

EMRs were the first electronic sources introduced with the aim of digitization patient information adding benefits, not present in paper charts, like the ability to easily collate and track sets of information, monitor changes in patient outcomes after implementation of a new practice or procedure, and determine which patients are due for physical exams, procedures, immunizations, and the like.

Unfortunately, EMRs are often practice-specific, making it difficult to transfer information to outside groups of providers, to other healthcare systems, and to patients. Because of such limitations, over time, EHRs were specifically designed for information sharing also between providers and patients. The introduction of this type of technology and the effort to make it more and more shareable seems to be an unavoidable trend.

A database with Electronic Health Records contains patient data recorded to varying levels of granularity. In particular, it contains:

- continuous physiologic monitoring data;
- laboratory measurements;
- medications administered;
- treatments and procedures;
- physician notes;
- imaging results.

In other words, advances in electronic medical record (EMR) technology have made it possible for the EMR to replace many functions of the traditional paper chart, and the use of EMR systems promises significant advances in patient care (Makoul et al. [2001]).

The presence of this new type of data has involved the development of research with the aim of using this data to support doctors' decisions.

Hodgson et al. [2019] observe that, within health care, clinical decision support systems (CDSS) are increasingly being introduced with the aim to provide pertinent information, intelligently filtered or presented at appropriate times, to enhance care and potentially improve outcomes.

Indeed, Electronic Health Records contain valuable data for identifying health outcomes, but these data also present numerous challenges. Statistics and Machine Learning methods could help with some of these challenges (Wong et al. [2018]).

Electronic health records are also considered as an instance of Real World Data (RWD). They are obtained outside the context of randomized controlled trials since generated during routine clinical practice.

The analysis of RWD is therefore primary to generating evidence about the real-world effects of medical products and health outcomes. In fact, EHR data may have utility for predicting health outcomes due to the presence of richer clinical information (including vital signs, laboratory test results, progress notes, and radiologic and pathologic images and reports).

The creation of computer-processable algorithms to identify individuals with specific health conditions, diseases, or provide a timely warning of risks of clinical events from electronic health data is often referred as 'computational phenotyping' (Wang [2020]).

In recent years the trend of adoption of digital health record systems in hospitals seems to be clear and no longer deferrable. In the US, for example, the number of hospitals with basic digital systems increased from 9.4% to 75.5 % over the period between 2008 and 2014 (Collins and Tabak [2014]).

Despite the progress realized in recent years, the EMRs and the EHRs data suffer yet of no standardization problem in measurements acquisition in the particular case of Intensive Care Unit (ICU).

Indeed interoperability of digital systems remains an open issue, leading to challenges in data integration. As a result, the potential that hospital data offers in terms of understanding and improving care is yet to be fully realized.

In parallel, the scientific research community is increasingly under criticism for the lack of reproducibility of studies (Collins and Tabak [2014]).

Recently some freely-available ICU databases have been made available for research purposes. In particular, MIMIC-III (Johnson et al. [2016]) is a large, freely-available database comprising de-identified health-related data associated with over 40000 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. The database includes information such as demographics, vital sign measurements made at the bedside, laboratory test results, procedures, medications, caregiver notes, imaging reports, and mortality. MIMIC supports a diverse range of analytic studies spanning epidemiology, clinical decision-rule improvement, and electronic tool development. MIMIC-III, by making it easier to access this type of data, has contributed to the dissemination of articles on the use of ICU data.

As highlighted by Shafaf and Malek [2019], the use of statistical methods as well as artificial intelligence and machine learning techniques in different medical fields are rapidly growing. In particular, studies belonged to three categories:

- prediction and early detection of disease;
- prediction of need for admission;
- discharge and mortality;

Different examples of models on disease prediction reported in Shafaf and Malek [2019] are applied on:

- AKI;
- influenza;
- sepsis;
- chronic obstructive pulmonary disease (COPD);
- urinary tract infection (UTI).

In Table 1.2 Shafaf and Malek [2019] list different models developed with the aim of disease prediction

As observed by Tsai et al. [2020] despite advances in the field of electronic health records (EHRs) implementation and adoption challenges persist, and the benefits realized remain below expectations

Disease	Algorithm	Result	Year
AKI	boosted ensembles of decision trees	AUC: 0.72- 0.87	2018 (8)
	Logistic regression	AUC: 0.77	2018 (9)
	Gradient Boosting Machine	AUC: 0.73-0.97	2018 (12)
	Deep Learning	Accuracy: 99.1%	2018 (13)
	Logistic regression	AUC : 0.74	2016 (10)
	Binary logistic regression	Sensitivity: 96.6% Specificity: 95.7%	2008 (11)
Influenza	Bayesian classifier (naïve Bayes)	AUC: 0.92-0.93	2015 (15)
	Bayesian network classifiers	AUC: 0.79	2014 (14)
Sepsis	Gradient tree boosting	AUC: 0.87-0.92	2018 (18)
	Support Vector Machine	AUC: 0.86	2017 (19)
COPD and Asthma	Random forest	C-statistic: 0.84	2018 (20)
	Logistic regression	Accuracy: 89.1%	2017 (21)
	Naive Bayes	Accuracy: 70.7%	2013 (23)
	Tree-based decision model	AUC: 0.83	2010 (22)
UTI	Extreme gradient boosting	AUC: 0.90	2018 (17)
Appendicitis	Rule base	AP: 0.86	2013 (25)

AKI: acute kidney injury; COPD: chronic obstructive pulmonary disease;
UTI: urinary tract infection; AP: average precision and recall.

Table 1.2: Models on disease prediction as reported in Shafaf and Malek [2019]

although positive effects of EHR implementation were identified, relating to clinical work, data, and information, patient care, and economic impact.

In the same review, Tsai et al. [2020] insert also information about the country of origin of the studies. The majority of the 141 studies analyzed by them are developed in the USA (81 studies) while only 6 are developed in a European country (1 in Italy).

1.3 Literature review on predicting Acute Kidney Injury using EMR

As previously mentioned, the advances in clinical informatics and the increasing availability of electronic medical records (EMR) have eased the development of predictive models of risk estimation in AKI.

Management of AKI presents a significant challenge and clinicians may be helped by robust prediction models for risk evaluation, foster prevention, and recognition. Electronic interventions also hold promise to systematically improve overall AKI care across healthcare settings from risk recognition to early detection and subsequent monitoring and management (Hodgson et al. [2019]).

In this paragraph, we insert a brief review of the literature on the methods applied to predict the onset of AKI. There is expanding literature on the problem, especially in recent years, demonstrating the growing interest in the topic published in almost all medical journals. For this reason, my overview is limited to a subset of the most recently published articles.

First of all, we start by mentioning the article of Gameiro et al. [2020] where a review of the literature on AKI risk prediction is carried out.

A summary of the studies is synthesized in the Table in 1.3.

In Gameiro et al. [2020], it is noted that almost all of these studies share common features:

Study	Design	Setting	N	AKI Definition	Timing of AKI	ML Algorithm	Predictive Ability
Kate *et al.* (2016)	retrospective	medical and surgical	25,521	AKIN	during hospitalization	naïve Bayes; support vector machine; decision trees; logistic regression	AUC 0.654 AUC 0.621 AUC 0.639 AUC 0.660
Thottakkara *et al.* (2016)	retrospective	surgical	50,318	KDIGO	post surgery	naïve Bayes; generalized additive model; logistic regression; support vector machine	AUC 0.819 AUC 0.858 AUC 0.853 AUC 0.857
Davis *et al.* (2017)	retrospective	medical and surgical	2003	KDIGO	during hospitalization	random forest; neural network; naïve Bayes; logistic regression	AUC 0.730 AUC 0.720 AUC 0.690 AUC 0.780
Cheng *et al.* (2018)	retrospective	medical and surgical	60,534	KDIGO, AKIN, RIFLE	during hospitalization	random forest; AdaBoostM1; logistic regression	AUC 0.765 AUC 0.75 AUC 0.763
Ibrahim *et al.* (2018)	prospective	contrast	889	KDIGO	pre and post intervention	logistic regression	AUC 0.790
Koola *et al.* (2018)	retrospective	medical and surgical	504	KDIGO	during hospitalization	logistic regression; naïve Bayes; support vector machines; random forest; gradient boosting	AUC 0.930 AUC 0.730 AUC 0.900 AUC 0.910 AUC 0.880
Lin *et al.* (2018)	retrospective	ICU	19,044	KDIGO	during hospitalization	support vector machine	AUC 0.860
Koyner *et al.* (2018)	retrospective	medical and surgical	121,158	KDIGO	24 h post admission	gradient boosting	AUC 0.900
Huang *et al.* (2018)	retrospective	PCI	947,091	AKIN	during hospitalization	gradient boost; logistic regression	AUC 0.728 AUC 0.717
Huang *et al.* (2019)	retrospective	PCI	2,076,69	AKIN	pre and post intervention	generalized additive model	AUC 0.777
Tomašev *et al.* (2019)	retrospective	medical and surgical	703,782	KDIGO	during hospitalization	recurrent neural network	AUC 0.921
Adhikari *et al.* (2019)	retrospective	surgical	2901	KDIGO	post surgery	random forest	AUC 0.860
Flechet *et al.* (2019)	prospective	ICU	252	KDIGO	during hospitalization	random forest	AUC 0.780
Parreco *et al.* (2019)	retrospective	medical and surgical	151,098	KDIGO	during hospitalization	gradient boosting; logistic regression; deep learning	AUC 0.834 AUC 0.827 AUC 0.817
Xu *et al.* (2019)	retrospective	medical and surgical	58,976	KDIGO	during hospitalization	gradient boosting	AUC 0.749
Tran *et al.* (2019)	prospective	burn	50	KDIGO	during hospitalization	k-nearest neighbor	AUC 0.920
Zhang *et al.* (2019)	retrospective	ICU	6682	KDIGO	24 h post admission	gradient boosting	AUC 0.860
Zimmerman *et al.* (2019)	retrospective	ICU	46,000	KDIGO	72 h post admission	logistic regression; random forest; neural network	AUC 0.783 AUC 0.779 AUC 0.796
Rashidi *et al.* (2020)	retrospective and prospective	burn and trauma	50/51	KDIGO vs New Biomarkers	1st week post ICU admission	recurrent neural network	AUC 0.920

AKI-acute kidney injury, AKIN-acute kidney injury network, AUC-area under the receiver operating characteristic curve, ICU-intensive care unit, KDIGO-kidney disease improving global outcomes, ML-machine learning, PCI-percutaneous coronary intervention, RIFLE-risk, injury, failure, loss of kidney function, end-stage kidney disease.

Table 1.3: Table with a review of results from Gameiro et al. [2020]

- use the definition of AKI KDIGO and as primary output the binary classification AKI-NO AKI;
- are retrospective and largely without external validation (although prospective studies are starting to be performed);
- apply a variety of supervised classification techniques in the same study (gradient boosting, logistic regression, k-nearest neighbor, Random Forest, deep learning, support vector machine,

naive Bayes, generalized additive model, recurrent neural network);

- have performances reported indicating the AUROC which for the various methods and in the various studies is variable between 0.7 and 0.93

Analogous methods are also reported in Huang et al. [2020], another article with a review of five studies (with partial overlap with the previous review) carried out on adult patients. Again, the studies use the AKI classification according to the KDIGO criterion (with and without using the Urine Output). The techniques cited were also in this case: logistic regression, and random forest.

The objective output of the supervised learning in these studies varies and includes, for example, the prediction of AKI within one day of admission to intensive care, within 48 hours of admission, and up to one day before the onset of the phenomenon.

The study reported in Simonov et al. [2019] focuses on data acquired from 2012 to 2016 of three hospitals using the same electronic medical record (avoiding in this way the standardization problems of EMRs). As an output, in addition to the primary one of AKI-NO AKI, the prediction of the onset of dialysis, death, and "sustained AKI" (defined as at least two consecutive creatinine values consistent with a diagnosis of AKI) are also evaluated.

This article is interesting for the careful description of the covariates involved in the analysis (demographics, vital signs, and laboratory data were obtained directly from the EHR) and the approach that assumes that the measures varying over time are assumed as constants until remeasured. In this case, logistic regression is used for binary classification.

Moving from reviews to individual studies an interesting research is presented in Flechet et al. [2017]. In this case, a model is developed and made usable with an online application (available here http://www.akipredictor.com/). It predicts AKI according to the KDIGO classification in adult patients without using the Urine Output. The model was elaborated using about 4000 patients extracted from the multicenter EPaNIC database. In this case, the objectives of the study were:

- predict the development of AKI in any of the stages during the first week of stay in intensive care;
- predict the development of stages 2 and 3 (AKI23) in the first week of stay in the ICU.

To achieve the previous objectives, 4 models have been developed:

- *Baseline model* that uses only demographic data known upon admission;
- *Admission model* which uses, in addition to the previous ones, the data available for admission;
- *Day1 model* in which the data available on the first day of stay in therapy are added;
- *Day1 + model* which uses additional data acquired in the week prior to admission to intensive care.

The selection of the variables was carried out starting from the univariate association of the candidate variables to be considered as predictors of AKI. The Random Forest technique was applied to the subset of the selected variables.

A subsequent publication of the same authors Flechet et al. [2019], it is shown a comparison between the ability to predict the onset of AKI using the AKIpredictor model with the prediction

of doctors. The study was carried out prospectively and showed similar capabilities of the model compared to doctors who, however, according to the study, tend to overestimate the risk of AKI. This article is reported because it is a prospective study and shows an interesting comparison between the "algorithm" and the clinicians' estimate.

To our knowledge, Tomašev et al. [2019] is the most extensive study on the subject of continuous AKI prediction. It was carried out on the data of 703782 patients between 18 and 99 years old collected by the US Department of Veterans Affairs. Also in this case the classification according to the KDIGO criterion was used as the ground truth label (without using the Urine Output, not present in the majority of the data). Also, in this case, a deep learning algorithm has been developed that can predict the onset of AKI in a subsequent time window from 6-h to 72-h (in particular 48h). The event is considered positive if the onset of the AKI of the stage of interest occurs in that window (in particular the option at any stage), and otherwise negative. The measures varying over time of each patient were "discretized" in time windows of 6 hours (4 events per day of hospitalization). In this case, deep learning techniques were applied to carry out the classification, in particular, a Recurrent Neural Network was used. According to what was reported, the model was able to predict 55.8 % of AKI episodes and 90.2 % of AKI episodes requiring subsequent dialysis.

Lei et al. [2020] is a particularly relevant publication in that the characteristics of this study are similar to the type of data that will be subsequently analyzed by us. This retrospective study, in fact, is aimed at predicting the onset of AKI in 897 patients undergoing cardiac surgery within the first 7 days of hospitalization (although this study limits to a single type of intervention: aortic arch surgery). Also in this case the KDIGO criterion is used for the classification of AKI without using the UO because, according to the authors, it was a measure acquired in an unreliable way in the data in their possession. The following techniques were used by applying both a binary and a multiclass classification to the various stages of AKI: logistic regression, random forest, support vector machine, and Light gradient machine. According to the authors, the best results were obtained using the Light gradient machine with an AUC-ROC of 0.8 (binary classification case).

The article published by Morid et al. [2019] is characterized by the idea of inscribing the AKI problem in the broader field of the prediction of adverse events in intensive care and for the discussion of possible approaches to managing the time series of data. The application example shown in the article is created using data from the MIMIC-III database and having as its purpose the prediction of the AKI using the KDIGO criterion (without UO); the problem of managing the time series of data and the subsequent classification is seen in detail and tackled as a "pattern recognition" problem. In other words, the article describes the different approaches to "transform" a time series into "features" that can subsequently be used as input for the classification algorithms. The approaches described are for each patient's time series:

- Static transformation: the series is replaced by a representative value (e.g. average, first value, last value, most recent value);
- Dynamic transformation: the series is split into a series of consecutive time windows of fixed intervals not overlapping. Subsequently, for each of these time windows are identified the characteristic patterns.

In the article, in addition to the local trend of each time window, a global trend is also extrapolated that takes into account the global trend, for example, of the averages of the different time windows.

Once the series of features have been built, the study applies many statistical techniques to it (logistic regression, random forest, kernel-based Bayesian Network, SVM, neural networks, and gradient boosting tree). Among these, the best result of the study is obtained with the use of random forests.

In a more recent publication by Song et al. [2021] a review of the most recent article focused on predicting AKI makes a comparison between the results obtained using machine learning and logistic regression techniques. According to Song et al. [2021], although machine learning and logistic regression have Area under the curve (AUC) similar however ML models exhibited variable performance with some ML displaying very good performance.

In Song et al. [2021] is reported a scatter plot of model performance comparing logistic regression and ML models (see 1.1).

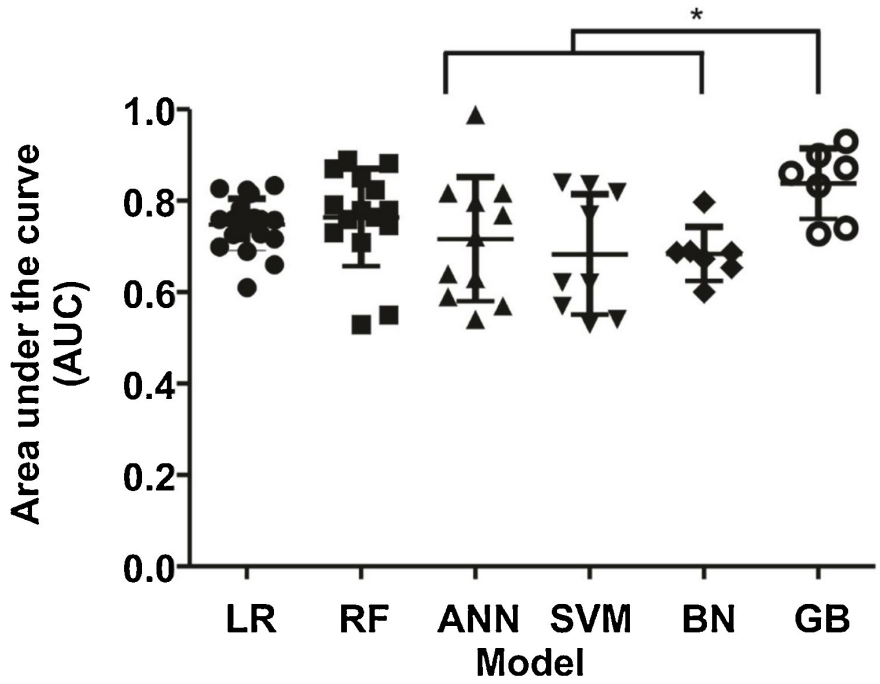

Figure 1.1: Scatter plot of model performance comparing logistic regression (LR), Gradient boosting (GB), artificial neural network (ANN), support vector machine (SVM), Bayesian Network (BN).

Gradient boosting with the exception of the random forest was significantly more effective at predicting AKI compared with ANN, SVM, and BN. Song et al. [2021] refer that although significant predictors were unique and study specific, the analyzed studies the creatinine results be the most common predictor.

A final mention to the article of Sanchez-Pinto and Khemani [2016] is due because it was carried out on the data reported from the electronic records of a pediatric intensive care unit (even if not cardiac) by analyzing patients from 1 month to 21 years of age. The goal of the study is to predict the onset of AKI at any stage within 72 hours of admission using data from the first 12 hours. Also in this case the KDIGO criterion without UO is chosen as a reference for the classification.

In the end, we report that studies that use EMRs in order to predict the AKI stage focused on pediatric intensive care are rare and no studies, at least to the best of our knowledge, have been performed in pediatric cardiac patients admitted to the intensive care unit.

Chapter 2

EMRs data from "Ospedale Pediatrico Bambino Gesù"

Ospedale Pediatrico Bambino Gesù (OPBG) is the leading children's hospital in Italy, being a key center for the health of children and adolescents from both Italy and abroad. It was founded in the second half of the 19^{th} century and today it is the largest sub-specialty children's hospital and research center in Europe. OPBG collaborates with the main international organizations in the pediatric field and treats a large number of patients. The hospital covers all pediatric sub-specialties: transplantation, genetic and metabolic diseases, neuroscience and rehabilitation, and, in particular, cardiology and cardiac surgery.

Children's Hospital Bambino Gesù is recognized at the national level as an Institute for research, hospitalization, and health care in pediatric patients. OPBG is also accredited as Academic Hospital by Joint Commission International, a non-governmental, non-profit organization that has been dedicated to improving the quality, safety, and efficiency of healthcare for over 75 years.

In particular, among the various departments of the hospital, there is the Pediatric Cardiac Intensive Care Unit (PCICU). It is committed to caring for all patients with acquired and congenital heart disease from fetal life to adulthood. The PCICU's team is composed of cardiac intensivists, cardiac surgeons, cardiologists, cardiac anesthesiologists, and nurses who collaborate in a state-of-the-art facility to provide high-quality care. A significant part of Italian children with congenital heart disease is treated in PCICU of OPBG.

Starting from the end of 2016, the $EMRs$ were introduced in the intensive care unit and subsequently, a project has been approved with the aim to use retrospectively this new type of EMR data collected at the OPBG PCICU to the problem of AKI.

In this chapter, we present the characteristics of the EMRs of Pediatric Cardiac Intensive Care of Children's Hospital "Bambino Gesù" and describe the process that led to the selection and extraction of the information present in the database. In the absence of an established data standardization practice for the treatment of the AKI problem, this process lasted several months involving many meetings with doctors and IT technicians.

Our retrospective study involved patient records extracted from January 2018 to February 2020. All the data extracted by the EMR have been anonymized. The ultimate goal is to develop a system capable of performing a continuous prediction of the AKI status for each patient with such an advance in time to allow doctors to decide on any changes to therapies.

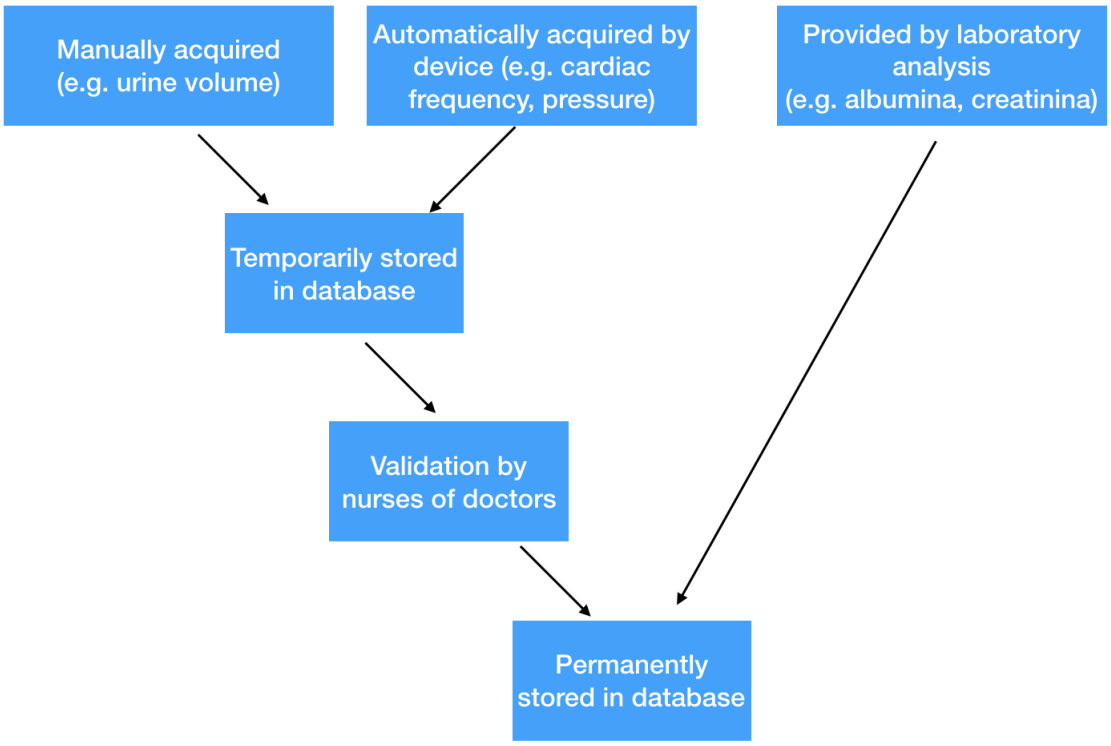

Figure 2.1: Database schema

2.1 Characteristics of OPBG PCICU EMRs

At the end of 2016, PCICU introduced the Ascom Digistat a comprehensive suite of clinical workflow software provided by Ascom (https://www.ascom.com). After a period of experimentation, this software suite has become part of PICU's clinical practice. After this initial phase, it was possible to collect all patient data regarding in particular:

- continuous physiologic monitoring data;
- laboratory measurements;
- treatments and procedures administered;
- blood gas analysis;
- admission data and doctor's summary notes.

It is important to notice that although the continuous physiologic monitoring data are acquired with an acquisition time of the order of the second, the database is designed to store permanently only the subset of these acquisitions validated by nurses/doctors as described in figure 2.1

This choice allows a better quality data decreasing the number of spurious measurements and storing only values that are compatible with the patient's status on the basis of the opinion of those working in the ICU. In particular, this choice largely avoids the acquisition made with sensors not correctly positioned due to patient movements.

The choice of the variables involved was made following doctors' indication taking into account the literature on AKI and the specific characteristics of the pediatric intensive care unit.

Type of variable	name of variable
Admission and post-surgical data (fixed data)	sex, age, weight, PIM3, Vasoactive-inotropic Score (VIScore), surgical state, baseline creatinine, CEC Duration, Clamping Duration
Vital Signs	systolic pressure, diastolic pressure, average pressure, saturation, cardiac frequency
Fluids	diuresis, fluid input, fluid output, blood input, blood output
Blood gas analysis	Be, Na+, Cl-, Lac, blood PH
Laboratory analysis	creatinine, albumin, hemoglobin, platelets, LDH, APPT
Therapies administered	Adrenaline, Milrinone, Furosemide (Lasix), Levosimendan, Vasopressin, Etacrynic acid

Table 2.2: Selected PCICU variables

We use a selection of objectively collected markers through the available EMR system. This information can be grouped according to their characteristics into groups such as patient demographics, vital parameters, laboratory analysis, blood gas analysis, etc...

Some of these variables are static and not subject to change during hospitalization (e.g. age or sex) whereas others might be susceptible to substantial changes over time (eg. pressure, saturation, diuresis, cardiac frequency, etc).

We include all such features that are collected both manually (such as urinary volume) or automatically (such as cardiac frequency, pressure, saturation) from patients' monitors and devices.

In our study, we use a subset of measures reported in Table 2.2 selected following the PCICU doctors' indications.

Currently, the data stored in the database are only those validated by nurses or doctors or provided by blood gas analysis or laboratory data. For this reason, different measures have different acquisition frequencies as described in Table 2.3.

Mean acquisition time means that for medical reasons a measurement could be acquired more frequently (for example laboratory analysis of blood gas analysis) by doctors and if non requested it is assumed to be stable in time.

Type of variable	Mean Acquistion Time
Vitals Signs	$\sim$ 2 h
Fluids	$\sim$ 2 h
Blood gas Analysis	$\sim$ 4 h
Laboratory analysis	$\sim$ 24 h

Table 2.3: Mean acquisition time for each group of variables.

Regarding the therapies administered, they are inserted in the database according to some conventional rules involving the start time and assumed duration and the administered dose in each

time interval must be computed.

It is important to notice that a pediatric patient admitted to intensive care can be subjected, although not frequently, to more than one surgical and/or hemodynamic procedure during the same hospitalization.

Due to the complexity of the data acquired on a patient in the ICU (including not standardized data management or the lack of data between the successive procedure) and according to PCICU doctors we decided to select a subset of the dataset. In particular, we use:

- only patient in pediatric age ($\leq$ 18 years) with a length of hospitalization greater than 48h;
- only the temporal data between admission and discharge date from PCICU or between the end time of the first surgery (or first cardiac procedure) and the start time of subsequent surgery.

2.1.1 Admission and post-surgical data

In this section, we describe the variables that are acquired in PCICU admission phase or obtained after the surgical or hemodynamic procedures. In particular, we notice that:

- pediatric Index of mortality (PIM3) is a severity scoring system used for predicting the outcome of patients admitted to pediatric intensive care units based on data collected within the first hour of admission;
- vasoactive-inotropic Score (VIS score) is a measure of post-operative cardiovascular support, and has been associated with morbidity and mortality after infant cardiac surgery;
- baseline creatinine is the value of serum creatinine which is reflective of the patient's premorbid kidney function;
- extracorporeal circulation (CEC) is used to maintain the patient's blood circulation and/or lung function outside the body. The CEC duration is related to the difficulty of the surgical procedure and the consequent risk;
- clamping duration indicates, instead, the length of clamping the aorta and separates the systemic circulation from the outflow of the heart.

2.1.2 Vital signs and fluids

Nurses and clinicians have traditionally relied on many vital signs to assess their patients. In the particular case of a cardiac intensive care unit, the selected variables are pressure (systolic, diastolic, and average), heart rate, and saturation.

It is important to notice that we use the invasive measurement of continuous arterial pressure except in case a non-invasive contemporary measure is stored in the database.

Although invasive measurement from an arterial line is generally considered to be the gold standard, for many practical reasons (e.g. movement artifacts) nurses and doctors proceed with the acquisition of non-invasive measurement when the invasive one is considered not reliable. Concerning the group of variables denoted in Table 2.2 as fluids, these variables are linked related to fluid and blood management. All the fluids in input and output (non only the diuresis) are strictly controlled. The same approach is obviously used for blood loss or transfusions.

2.1.3 Blood gas analysis

Blood gas analysis is a common investigation used to assess and monitor the acid–base balance of patients. Blood gas analysis can inform healthcare professionals about the respiratory and metabolic status of their critically ill patients. Concerning blood gas analysis we use the following variables:

- base excess (BE) or base deficit is characterized by the amount of base that is required to normalize the pH of the blood;
- Na+, Cl-, Lac indicate the concentration for potassium (Na), chloride (Cl), and lactate (Lac);
- blood PH indicates the pH of the blood and refers to how acidic it is.

2.1.4 Laboratory analyses

Laboratory testing is ubiquitous among hospitalized patients and is more common among patients in the intensive care unit (ICU). The execution of laboratory tests is frequently requested for the diagnosis and/or monitoring of critical patients.

The variables included in the group of laboratory analyses are:

- creatinine, a product made by muscles as part of regular, everyday activity. Normally, kidneys filter creatinine from your blood and send it out of the body in urine. If there is a problem with the kidneys, creatinine can build up in the blood and less will be released in the urine;
- albumin is a plasma protein; low albumin levels can be a sign of liver or kidney disease or another medical condition. High levels may be a sign of dehydration;
- hemoglobin is the protein contained in red blood cells that is responsible for the delivery of oxygen to the tissues and transports carbon dioxide from organs and tissues back to the lungs;
- platelets are a component of blood whose function (along with the coagulation factors) is to react to bleeding from blood vessel injury by clumping, thereby initiating a blood clot;
- lactate dehydrogenase (LDH) is an enzyme found in nearly all living cells. Because it is released during tissue damage, it is a marker of common injuries and diseases such as heart failure;
- a partial thromboplastin time (aPTT) measures the time it takes for a blood clot to form.

2.2 Time dependent data extraction

As highlighted in the introduction of this chapter, the final aim is to develop a system able to support doctors' decisions on any changes in therapies to prevent AKI. To allow this, a continuous forecasting approach of the state of AKI was chosen with a time advance (e.g. 48 hours) capable of making a forecast for each pre-selected time interval, possibly allowing the doctors to decide on any modifications to the therapies.

For this reason, according to PCICU clinicians, we use the approach to discretize all the different acquisition frequencies in a common sample frequency of $\Delta t = 6$ hours. In particular, in the case

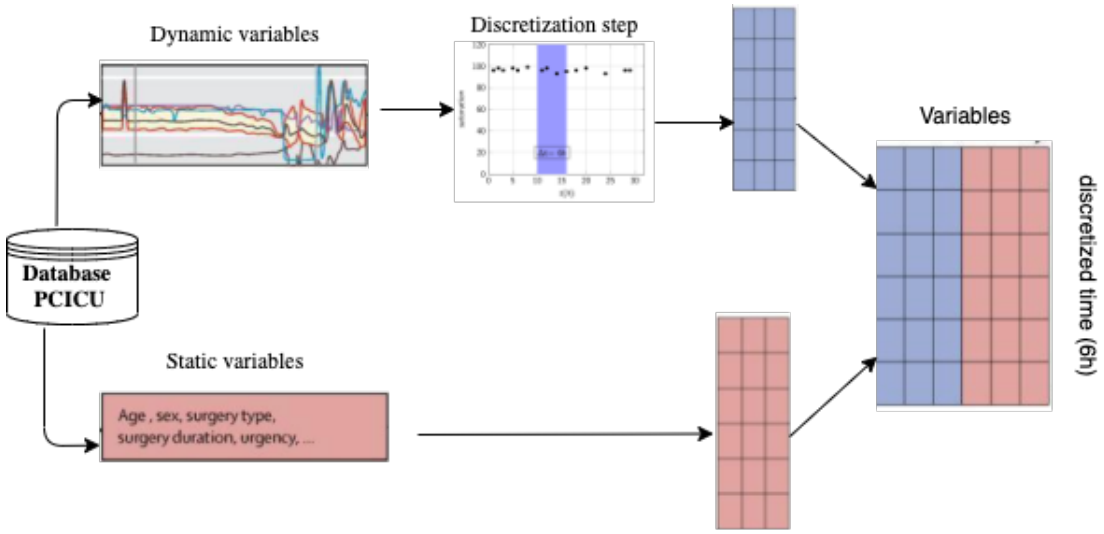

Figure 2.2: Data extrapolation schema

of measure acquired more frequently than 6 hours, we use, starting from the initial time, for each time interval Δt:

- the mean of data acquired in the case of intensive measures (e.g. pressure, cardiac frequency)
- the sum of data acquired in case of extensive measures (e.g. diuresis, blood output).

For the laboratory analysis case, instead, following the indications of the ICU doctors, we assume to repeat the most recent measure until new acquisition.

In addition to these data, the EMR records the AKI KIDIGO value in each hour of stay in the ICU.

In this case, we adopt different methods of "discretization". For each Δt we compute:

- bin AKI: 0 if all hourly records are zero, 1 otherwise;
- severe AKI: 0 if the values are all 0 or 1, 1 otherwise;
- max AKI: the maximum values assumed in each interval Δt;
- mode AKI: the mode in each interval Δt.

The first two values are binary values, while the other values assume values 0,1,2,3 following the AKI stage criteria.

The main objective of this study is to predict the stage of AKI (in particular bin AKI, severe AKI, max AKI, and mode AKI) that will occur after a certain temporal delay. This approach can be summarized using this schema:

$$6h\ data \rightarrow 48h\ temporal\ delay \rightarrow 6h\ data$$

In Figure 2.2 we depict the data extrapolation schema taking into account time-dependent and time-independent variables.

2.3 Therapies data extraction

Our study considers also some therapies administered to the patients. Between all the possible drugs, according to doctors' expertise, we select the following subset of drugs reported in the following table 2.4.

Name of the drug (unit of measure)	type of intravenous administration
Ethacrynic Acid (mg/kg)	infusion and bolus
Adrenaline (mcg/kg)	infusion
Furosemide (mg/kg)	infusion and bolus
Levosimendan (mg/kg)	infusion
Milrinone (mg/kg)	infusion
Vasopressin (UI/kg)	infusion

Table 2.4: List of drugs extracted from the database

Intravenously administered drugs are given either as a "bolus" or an infusion over a period of many hours. Bolus administration is rarely used and is restricted to emergency situations where rapid pharmacological action of limited duration is required.

In our case, the bolus was considered only in the case of furosemide and ethacrynic acid.

In Table 2.5 we show an example of how the information is stored in PCICU's database in case of infusion, where start and end time of a drug are entered together with the Amount, Speed, Volume and Weight.

IDEpisode	Name	DateUTC	Action	Amount	Speed	Volume	Weight
XXXX	Levosimendan	YYYY-12-01 08:00:17	Start	1.3	2.08	50	9
XXXX	Levosimendan	YYYY-12-02 07:40:58	Start	2	0.8	50	9
XXXX	Levosimendan	YYYY-12-03 07:34:53	End	2	0.8	50	9

Table 2.5: Example of database in the infusion case

In table 2.6, instead, we show the data in the case of the bolus.

IDEpisode	Name	DateUTC	Amount	Weight
YYYY	Furosemide (bolo)	YYYY-05-23 14:42:36	3	2.8
YYYY	Furosemide (bolo)	YYYY-05-26 06:04:25	2	2.8
YYYY	Furosemide (bolo)	YYYY-05-26 17:31:20	2	2.8

Table 2.6: Example of the database in the bolus case

The database doesn't provide the actual hourly dose administered to the patient, but only the prescribed dose that, in some cases, changes with the time or terminates beforehand. Therefore, the dose was then calculated by us starting from the database data.

In the case of Furosemide and vasopressin, we consider also the bolus dose.

Since all previous variables were discretized every Δt (6 hours), the dose was calculated and administered to the patient from the initial time every Δt. In the event that the bolus is also present

in that drug, for each time interval of 6 hours, the dose final is given by the sum of the drug given as an infusion and bolus.

In the table, we show an example of the database with the therapeutic indication.

For the bolus case, the dose is evaluated with the following formula:

$$dose = \frac{Amount(mg)}{Weight(kg)}$$

.

In the infusion case, instead, the dose is calculated using the following relation:

$$dose = \frac{Amount(mg)}{Volume(ml) \times Weight(kg)} \times Speed(ml/h) \times \Delta t(h)$$

with Δt the chosen time interval, Weight the weight of the patient, and Speed the speed of intravenous infusion.

2.4 Missing data

Some variables involved in our study are affected by the problem of missing values. For variables belonging to groups of fluids and therapies, there are no missing data because, when a value is not present that means that it is zero.

Concerning the Laboratory analysis, the clinical practice usually assumes to make one measurement for days or more frequently based on the patient's condition and the doctor's instructions. If laboratory analyses are not requested it is possible to assume, following the indications of the PCICU doctors, that the values have not changed compared to the previous ones and for this reason could be repeated until a new acquisition is inserted in the database.

In the case of variables relating to vital parameters and blood gas analysis, we assume to have missing data if there is no acquisition in the average acquisition time interval (respectively 2 hours for vital signs and 4 hours for blood gas analysis).

In table 2.7 we report information about missing data in the case of vital signs. Instead in table 2.8 we insert the missing data in case of blood gas analysis.

Variable name	number of missing data	total number of acquisitions
systolic pressure	3654	64168
diastolic pressure	3300	64168
average pressure	6825	64168
heart rate	1567	64168
saturation	2970	64168

Table 2.7: Missing data in case of vital signs

As described in Rubin [1976], we assume the origin of the missing data is Missing at Random (MAR). Missing data are regarded to be MAR when the probability that the responses are missing depends on the set of observed responses but is not related to the specific missing values which are expected to be obtained.

Variable name	number of missing data	total number of acquisitions
Na+	5554	31642
Cl-	5716	31642
bloodPH	5514	31642
Lac	5552	31642
BE	5595	31642

Table 2.8: Missing data in case of blood gas analysis

Starting this assumption of MAR, in the case of Blood gas analysis and Vital Signs we use a nonparametric missing value imputation using Random Forest provided by MissForest R Package.

As described in Stekhoven and Bühlmann [2011] MissForest, allows for missing value imputation on basically any kind of data. In particular, it can handle multivariate data consisting of continuous and categorical variables simultaneously. Moreover, it handles tuning parameters and requires no assumptions about distributional aspects of the data.

MissForest is an iterative imputation method based on a random forest. This approach, averaging over many unpruned classification or regression trees, takes advantage of the fact that random forest intrinsically constitutes a multiple imputation scheme.

Given $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_p)$ variables and a $n \times p$ matrix of data, this method for an arbitrary variable $\mathbf{X}_s$ (which includes missing values) splits the dataset into to four parts:

- the observed values of $\mathbf{X}_s$ denoted by $\mathbf{y}_{obs}^{(s)}$;
- the missing of $\mathbf{X}_s$, denoted by $\mathbf{y}_{mis}^{(s)}$;
- the variables other than $\mathbf{X}_s$ with observations denoted by $\mathbf{x}_{obs}^{(s)}$;
- the variable other than $\mathbf{X}_s$ with missing observation denoted by $\mathbf{x}_{mis}^{(s)}$.

The pseudo Algorithm 1 depicted in figure 2.3 gives a representation of the missForest method.

The imputation has been computed separately for blood gas analysis and vital signs given the different mean times of measurement acquisition.

In particular, Stekhoven and Bühlmann [2011] shows on several real datasets coming from different biological and medical fields that MissForest outperforms established imputation methods like k-nearest neighbors imputation or multivariate imputation using chained equations.

2.5 Summary statistics of selected dataset

In this section we insert some brief descriptive statistics evaluated on the selected subset of patients with:

- pediatric age
- time of permanence in PCICU such as to allow a continuous prediction using the temporal delay of $48h$

Algorithm 1 Impute missing values with RF.

Require: $\mathbf{X}$ an $n \times p$ matrix, stopping criterion γ

1. Make initial guess for missing values;
2. $\mathbf{k} \leftarrow$ vector of sorted indices of columns in $\mathbf{X}$ w.r.t. increasing amount of missing values;
3. **while** not γ **do**
4. $\quad \mathbf{X}^{\text{imp}}_{\text{old}} \leftarrow$ store previously imputed matrix;
5. $\quad$ **for** s in $\mathbf{k}$ **do**
6. $\quad\quad$ Fit a random forest: $\mathbf{y}^{(s)}_{\text{obs}} \sim \mathbf{x}^{(s)}_{\text{obs}}$;
7. $\quad\quad$ Predict $\mathbf{y}^{(s)}_{\text{mis}}$ using $\mathbf{x}^{(s)}_{\text{mis}}$;
8. $\quad\quad \mathbf{X}^{\text{imp}}_{\text{new}} \leftarrow$ update imputed matrix, using predicted $\mathbf{y}^{(s)}_{\text{mis}}$;
9. $\quad$ **end for**
10. $\quad$ update γ.
11. **end while**
12. **return** the imputed matrix $\mathbf{X}^{\text{imp}}$

Figure 2.3: The pseudo Algorithm of the MissForest method

Following this choice, the total number of patients on which it was possible to perform an analysis is 419 (193 female, 226 male). These patients are divided by surgical state within the following groups: Surgical and Hemodynamic (48), Surgical (299), Hemodynamic (34), and Medical (38).

In table 2.9 we insert the summary statistics related to the variables deriving from admission and post-surgical data.

Variables	Min, Q_1 Median Q_3 Max; $\mathbf{N = 419}$
weight (kg)	1 3 6 12 97
age (month)	0 0 5 32 216
basal creatinine (mg/dL)	0.08 0.26 0.35 0.58 1.37
PIM3	0 0.01 0.02 0.04 1
Vis Score	0 0 8 15 50
CEC Duration (min)	0 0 85 160 470
clamping duration (min)	0 0 22 80 294

Table 2.9: Summary statistics of variables related to admission and post-surgical data

In table 2.10 the time-dependent summary statistics are reported.

Variables	Min, Q_1, Median, Q_3, Max; **N = 12,852**
diuresis (ml)	0 95 160 310 2,600
fluid input (ml)	3 104 158 306 2,569
fluid output (ml)	0 100 165 320 2,660
creatinine (mg/dl)	0.07 0.26 0.37 0.56 11.05
ldh (mU/ml)	143 351 479 706 6,389
albumin (g/dl)	2.00 3.70 4.10 4.70 7.50
hemoglobin (g/dl)	6.90 11.10 12.10 13.30 22.40
platelets ($\times 10^3/mm^3$)	14 132 226 341 901
aPTT (s)	20 31 36 43 178
BE (mEq/l)	-12.8 3.4 6.4 10.1 40.2
Cl- (mmol/l)	70.0 97.2 100.6 104.0
Lac (s)	0.40 1.00 1.23 1.60 18.57
Na+ (mmol/l)	118.0 135.5 138.0 141.4 161.0
blood PH	7.08 7.42 7.45 7.49 7.72
blood input (ml)	0 0 0 0 2,200
blood output (ml)	0 0 0 6 770
systolic press (mmHg)	42 76 88 100 150
diastolic press (mmHg)	16 42 52 61
average press (mmHg)	25 54 63 72 107
heart rate (bpm)	46 108 126 142 220
saturation	51.3 92.3 97.7 99.5 100.0
adrenaline (mcg/kg)	0 0 0 6 144
furosemide (mg/kg)	0.0 0.0 0.0 0.0 5.00
ethacrynic acid (mg/kg)	0.00 0.0 0.77 1.67 6.52
milrinone (mg/kg)	0.0 0.0 0.0 0.18 30.0
levosimendan (mg/kg)	0.000 0.000 0.000 0.000 0.076
vasopressin (UI/kg)	0.00 0.00, 0.00 0.00 0.48

Table 2.10: Summary statistics of time-dependent variables: Min, Max, median (IQR)

In figure 2.4 we show boxplots of admission and post-surgical data separated according to the onset of AKI. In particular, we indicate by "No AKI" a patient in whom the onset of AKI does not occur and with "AKI" a patient in whom disease onset occurs at any stage.

In figure 2.5 we show the box plots of admission and post-surgical data separated according to the onset of severe AKI (stage 2 ore 3) or not severe (stage 0 or 1).

Finally, in figure 2.6 we show the boxplots of admission and post-surgical data separated according to the maximum stage of observed AKI.

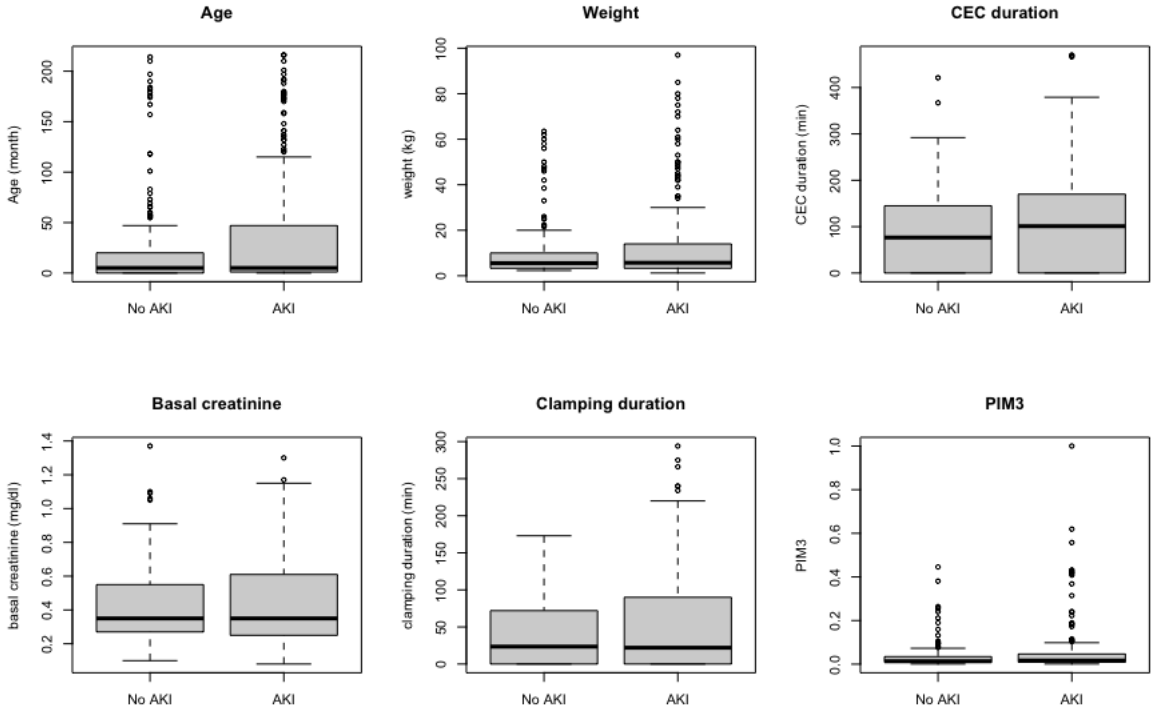

Figure 2.4: Box plots of admission and post-surgical data separated according to the onset of AKI or not.

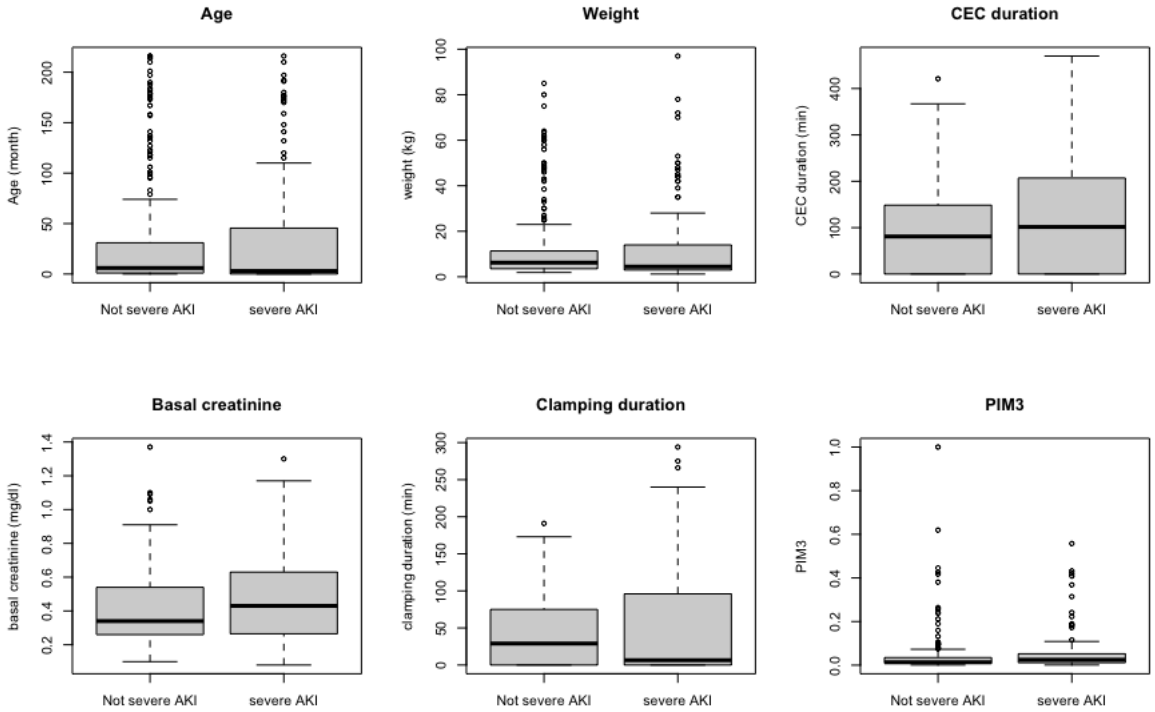

Figure 2.5: Boxplots of admission and post-surgical data separated according to the onset of AKI severe (stage 2 or 3) or not (no AKI or stage 1).

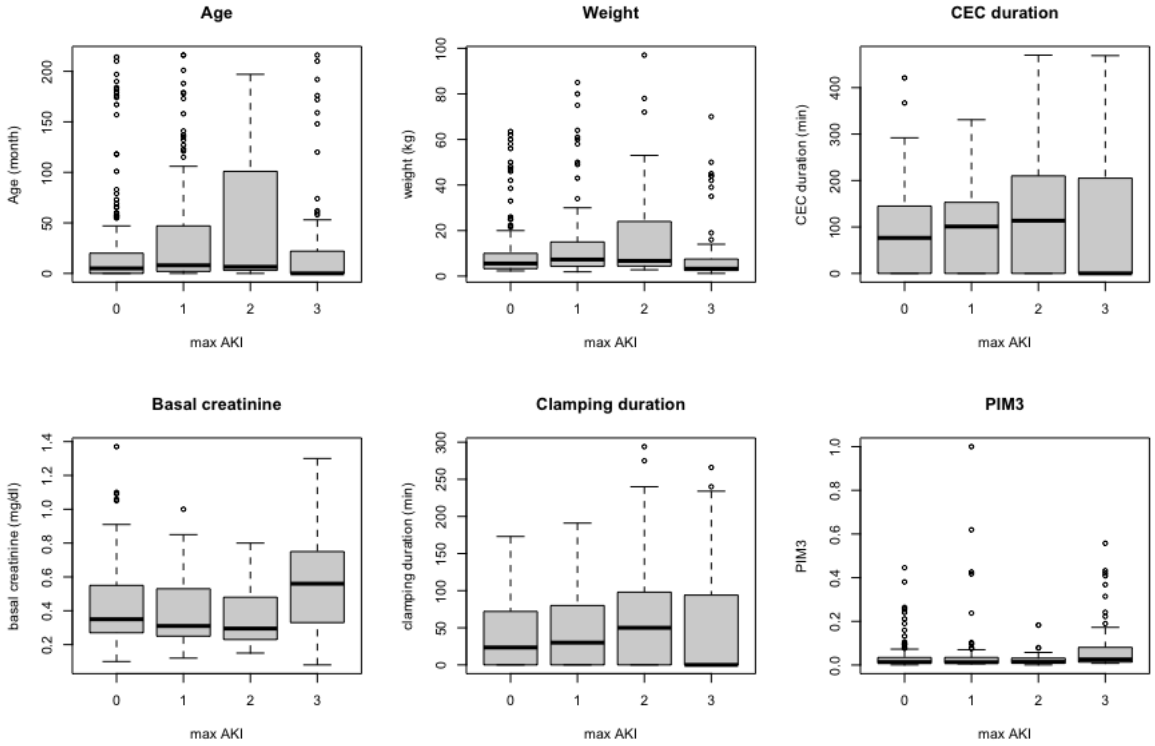

Figure 2.6: Boxplots of admission and post-surgical data separated by the maximum stage of observed AKI.

Chapter 3

Classification using Random Forest

3.1 Introduction

In this chapter, we show the results obtained applying Random Forests classification method to the PCICU dataset previously described in chapter 2. We have chosen this methodology after various preliminary attempts because it provides excellent performance and because, being an ensemble method that makes use of classification trees, it is characterized by being less black-box than other methods.

After a brief theoretical introduction to the random forest, we insert the classification results obtained by applying the RF model to:

- the selected subset of variables acquired in PCICU and the therapies administered;
- subsets of the previous variables identified using Recursive Feature Elimination (RFE) approach.

3.2 A brief introduction to Random Forest

The random forest algorithm was proposed by Breiman [2001]. It has been extremely successful as a general-purpose classification and regression method.

As highlighted by Biau and Scornet [2015], the popularity of forests depends on the fact that they can be applied to a wide range of prediction problems using a relatively small number of tuning parameters compared to other methods. The presence of a little number of parameters to tune makes using this algorithm easier than others.

Moreover, the method is generally recognized for its accuracy and its ability to deal with both small sample sizes and high-dimensional feature spaces. Another advantage is that it can be parallelizable.

Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest.

A pseudocode representing the random forest algorithm is displayed in Table 3.1.

As described by James et al. [2013] and Hastie et al. [2001] RF is an improvement over bagged trees by way of a random forest small tweak that decorrelates the trees. As in bagging, a number of decision trees on bootstrapped training samples are built. But when building these decision trees,

Random Forest for Regression or Classification.

1. For $b = 1$ to B:
   (a) Draw a bootstrap sample $\mathbf{Z}^*$ of size N from the training data.
   (b) Grow a random-forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum size n_{min} is reached.
      i. Select m variables at random from the p variables.
      ii. Pick the best variable/split-point among the m
      iii. Split the node into two daughters nodes.
2. Output the ensemble of trees T_{b1}^{B}

To make a prediction at a new point x:
Regression: $\hat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$
Classification: Let $\hat{C}_b(x)$ be the class prediction of the *bth* random-forest tree. Then $\hat{C}_{rf}(x) = majority\ vote\ \{\hat{C}_b(x)\}_1^B$

Table 3.1: Random Forest algorithm for regression or classification

each time a split in a tree is considered, a random sample of m predictors is chosen as split candidates from the full set of p predictors. The split is allowed to use only one of those m predictors.

A fresh sample of m predictors is taken at each split; typically it is chosen $m \approx \sqrt{p}$ where p the number of predictors considered at each split is approximately equal to the square root of the total number of predictors. This implies that, in the phase of building a random forest, at each split in the tree, the algorithm is not even allowed to consider a majority of the available predictors.

As suggested in James et al. [2013], intuitively this can be explained starting from the example of a case in which there is a strong predictor along with a number of other moderately strong predictors. Then in the collection of bagged trees, most or all of the trees will use this strong predictor in the top split. Consequently, all of the bagged trees will look quite similar to each other.

Hence the predictions from the bagged trees will be highly correlated and this unfortunately means that bagging will not lead to a substantial reduction in variance over a single tree in this setting. Random forests circumvent this problem by forcing each split to consider only a subset of the predictors.

Moreover, RF can be used to rank the importance of variables in regression or classification problems. The main method used to compute the variable importance makes use of the mean decrease in the Gini coefficient, a measure of how each variable contributes to the homogeneity of the nodes and leaves in the resulting random forest. The higher the value of mean decrease accuracy or mean decrease Gini score, the higher the importance of the variable in the model. Gini Importance or Mean Decrease in Impurity (MDI) (see Breiman [2001]) calculates each feature's importance as the sum over the number of splits (across all tress) that include the feature, proportionally to the number of samples it splits. The variable importance score is normalized by dividing all scores over the maximum score: the importance of the most important variable is always 100%.

Variable importance plots can be constructed for random forests in exactly the same way as

they were for gradient-boosted models. At each split in each tree, the improvement in the split criterion is the importance measure attributed to the splitting variable and is accumulated over all the trees in the forest separately for each variable.

On the theoretical side, several results have been obtained starting from Breiman [2001]; he demonstrates an upper bound on the generalization error of forests in terms of correlation and strength of the individual trees. Biau and Scornet [2015] present a review of several results obtained with the aim of narrowing the gap between theory and practice in particular on the consistency and asymptotic distribution of random forests.

3.3 Classification procedure and evaluation measures of the classification performance

In this section we describe the classification procedure implemented and the evaluation measures of classification performance used.

We implement a classification procedure using the `caret` (Classification And REgression Training) R package (see Max [2008] and Kuhn [2012]).

- We split the dataset in train (70%) and test (30%) sets. The former is used to fit the classification model, whereas the latter is employed to evaluate its performance;
- in splitting the data, we preserve the percentages of each class in train and test sets.
- in order to have more stable predictions and avoid the problem of overfitting, we use the k–Fold Cross–Validation (with $k = 10$) repeated 3 times;
- we perform tuning of the m (also noted as $mtry$) parameter using the random choice option in order to calibrate random forest parameters;
- finally, for this first model, we use the RF classification method.

In order to perform an evaluation measure of the classification performance we use the most common metric reported in the literature as highlighted in Chapter 1. All the results that will be shown in the next paragraphs are computed using the test dataset.

As described in Novakovic et al. [2017] and Labatut and Cherifi [2011] the evaluation of different classification models is a common task in which two cases are usually distinguished:

- binary classification with two possible classes;
- multiclass classification with more than two classes.

In the binary case, especially in the medical field, the most common measure is the Area Under the Curve (AUC) ROC, where ROC is the acronym of receiver operating characteristic curve.

As described in Fawcett [2006], the ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied taking into account also the unbalanced case. The method was originally developed for operators of military radar receivers

starting in 1941, which led to its name. In order to explain the AUC-ROC we start with the possible results of a classification with two classes. In binary classification, there are four possible outcomes for a test prediction: true positive (TP), false positive (FP), true negative (TN), and false negative (FN).

We also indicate with P and N respectively the number of positive and negative cases according to the True class. In figure 3.1 a classification scheme with two classes.

The true positive rate (TPR), or sensitivity, can be represented as:

$$TPR = \frac{TP}{TP + FN}$$

. In this case it is useful to use the ROC curve.
The true negative rate (TNR), or specificity, instead, is defined as

$$TNR = \frac{TN}{FP + TN}$$

. The false positive rate (FPR) is defined as:

$$FPR = \frac{FP}{FP + TN}$$

. FPR is also equal to:

$$FPR = 1 - specificity$$

. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate at various threshold settings. The true-positive rate is also known as sensitivity (or recall or probability of detection); the false-positive rate is also known as the probability of a false alarm and can be calculated as $1 - specificity$.

The Area Under the ROC curve (AUC-ROC) is a measure of how well a parameter can distinguish between two diagnostic groups.

As observed in Fawcett [2006], one point in ROC space is better than another if its TPR is higher, FPR is lower, or both and it is to the northwest of the first.

In multiclass cases, multiple metrics can be used as described in Grandini et al. [2020]. As reported also in Chapter , the most common metric is Accuracy computed starting from the confusion matrix.

The confusion matrix, a generalization of binary classification schema, is a cross table that records the number of occurrences between the true/actual classification and the predicted classification, as shown in Figure 3.2. The Accuracy is defined as:

$$Accuracy = \frac{number\ of\ correctly\ classified\ examples}{total\ number\ of\ cases}$$

Referring to the figure 3.2, the accuracy (or overall accuracy) is computed by the sum of diagonal elements on the confusion matrix divided by the total number of observations.

In binary classification, Accuracy is defined as:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

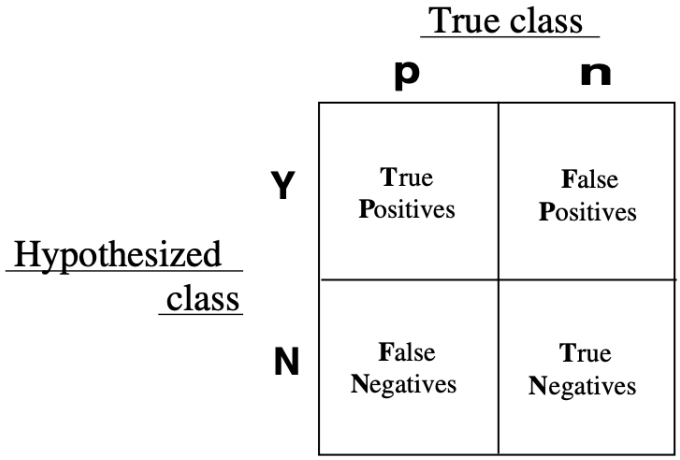

Figure 3.1: Binary classification schema (from Fawcett [2006])

Another useful measure is the so-called kappa index. Its origins are in the field of psychology with the aim of measuring the agreement between two human evaluators or raters (e.g., psychologists) when rating subjects (patients). It was later "appropriated" by the machine-learning community to measure inter-rater reliability for the qualitative (categorical) case.

As described in McHugh [2012], the Kappa result is interpreted as follows:

- values < 0.20 as indicating no agreement;
- 0.21–0.40 as fair;
- 0.41–0.60 as moderate;
- 0.61–0.80 as substantial;
- 0.81–1.00 as almost perfect agreement.

3.4 Classification results using all variables

In this section we show the results obtained by applying random forest using all the possible variables previously described in the following cases:

- binary AKI;
- severe AKI;
- maximum AKI;
- mode AKI.

All the results that will be shown below are computed using the test dataset not used in the training phase.

In all the cases we perform tuning of m parameter using the random choice available in the `caret` R package with the same tuneLength option equal to 15 and using a number of trees ntree = 500.

		PREDICTED classification				
	Classes	a	b	c	d	Total
ACTUAL classification	a	6	0	1	2	9
	b	3	9	1	1	14
	c	1	0	10	2	13
	d	1	2	1	12	16
	Total	11	11	13	17	**52**

Figure 3.2: Example of confusion matrix in multiclass case (from Grandini et al. [2020])

The processing time for training each model is about 3 hours with a MacBook Pro Intel Core i5 dual-core with 2.3 GHz and 16 GB of RAM. In `caret`, it is possible to use parallelization and the time can be reduced by exploiting the availability of a greater number of cores.

3.4.1 Binary and severe AKI case

In this paragraph, we show the results in binary and severe AKI cases. The results obtained in the binary AKI case are shown in Table 3.2. In particular, considering the reference class AKI 1, we obtain:

- $AUC - ROC = 0.93$
- $sensitivity = 0.71$
- $specificity = 0.99$

The sensitivity and specificity reported are computed using the typical $p = 0.5$ cut-off. More detailed information is reported in Appendix A.

	48h binary AKI score all variables	CI
SENS	0.71	0.68-0.74
SPEC	0.99	0.98-1
TP	523	-
FP	45	-
TN	3075	-
FN	212	-
AUC-ROC	0.93	0.92-0.94

Table 3.2: 48h classification results via Random Forest with therapies (binary AKI case)

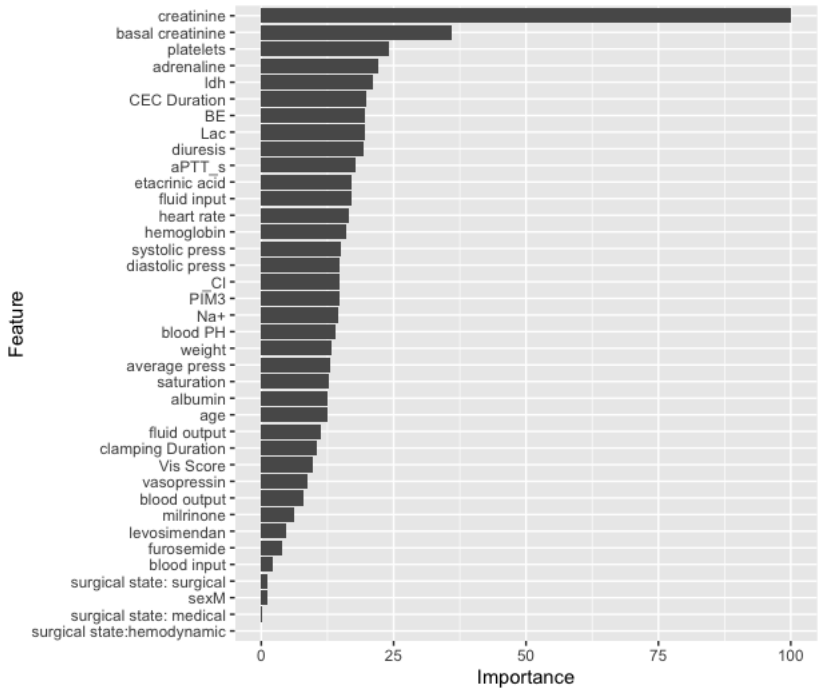

Figure 3.3: Variable importance case binary AKI

In figure 3.3 we insert the corresponding variable importance.

The results obtained in the severe AKI case are, instead, shown in Table 3.3. In this case, we have:

- $AUC - ROC = 0.99$
- $sensitivity = 0.74$
- $specificity = 0.994$

	48h severe AKI score all variables	CI
SENS	0.74	0.70-0.78
SPEC	0.99	0.99-1
TP	350	-
FP	20	-
TN	3362	-
FN	123	-
AUC-ROC	0.99	0.98-1

Table 3.3: 48h classification results via Random Forest with therapies (severe AKI case)

The results of AUC-ROC in the case of binary and severe AKI are very good compared, in particular, with results presented in the literature and reported in Chapter 1.

3.4.2 Maximum and Mode AKI case

In this paragraph, we show the results obtained in the multiclass case. We report the confusion matrix achieved and the corresponding values of accuracy and kappa. In Table 3.4 is reported the

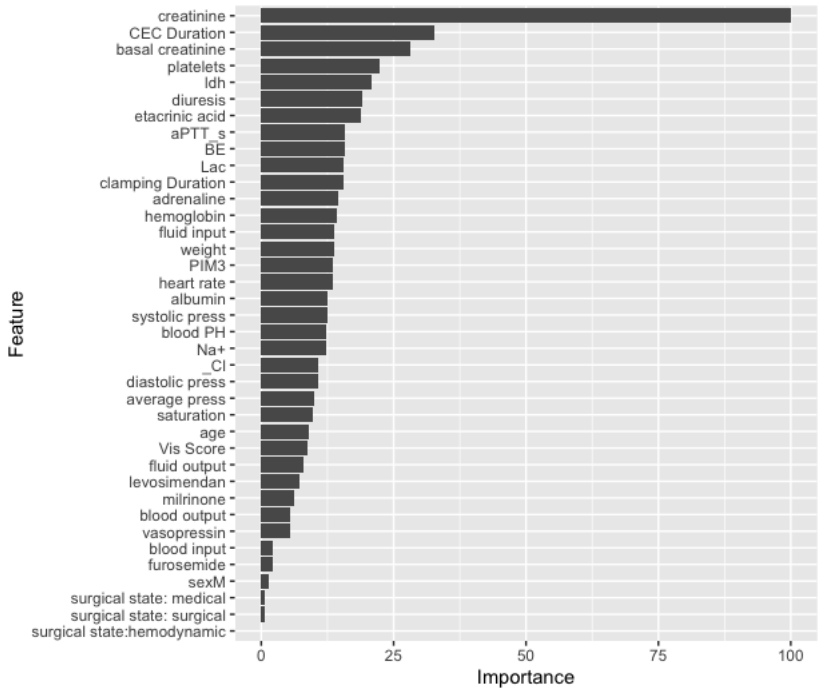

Figure 3.4: Variable importance case severe AKI

Max AKI confusion matrix				
	Observed			
Prediction	0	1	2	3
0	3084	165	52	36
1	11	88	5	7
2	7	6	123	4
3	18	2	6	239

Table 3.4: Confusion matrix: max AKI case using all variables

confusion matrix obtained in the maximum AKI case. The overall accuracy, in this case, is the following:

- Accuracy: 0.92 (CI: 0.91, 0.93)
- Kappa: 0.71

The variable importance for the maximum AKI case is depicted in Figure 3.5.

More detailed information is reported in A.

In Table 3.5 is reported the confusion matrix obtained in the mode AKI case. The accuracy for the mode AKI case is the following:

- Accuracy: 0.95 (CI: 0.94, 0.96)
- Kappa: 0.80

The variable importance for the mode AKI case is depicted in Figure 3.6. More detailed information is reported in A.

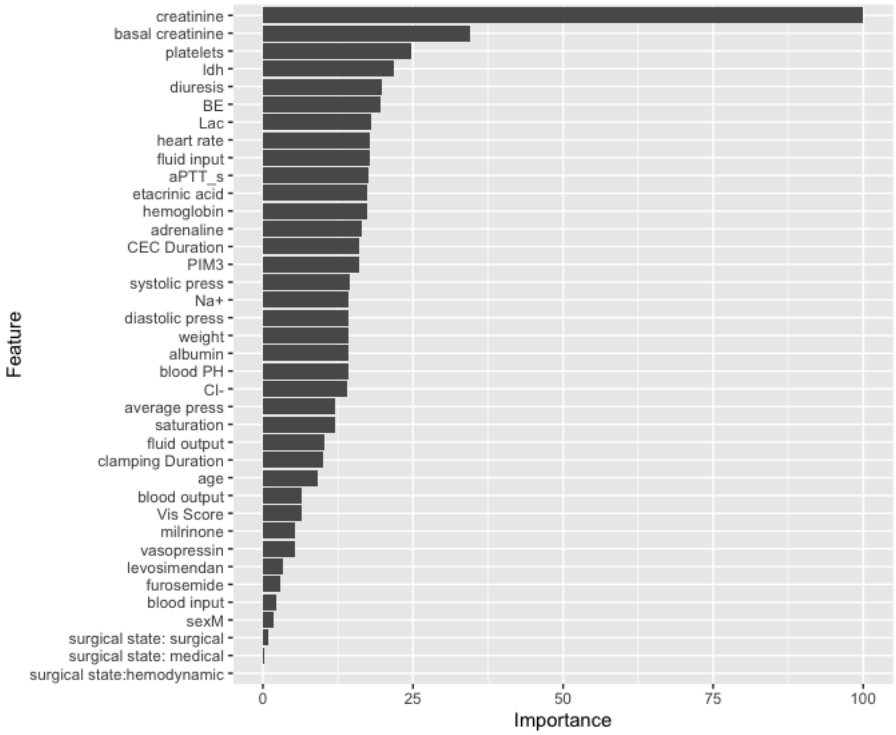

Figure 3.5: Variable importance case maximum AKI

Mode AKI confusion matrix				
	Observed			
Prediction	0	1	2	3
0	3232	83	26	43
1	6	76	4	0
2	2	4	136	4
3	15	4	6	213

Table 3.5: Confusion matrix: mode AKI case using all variables

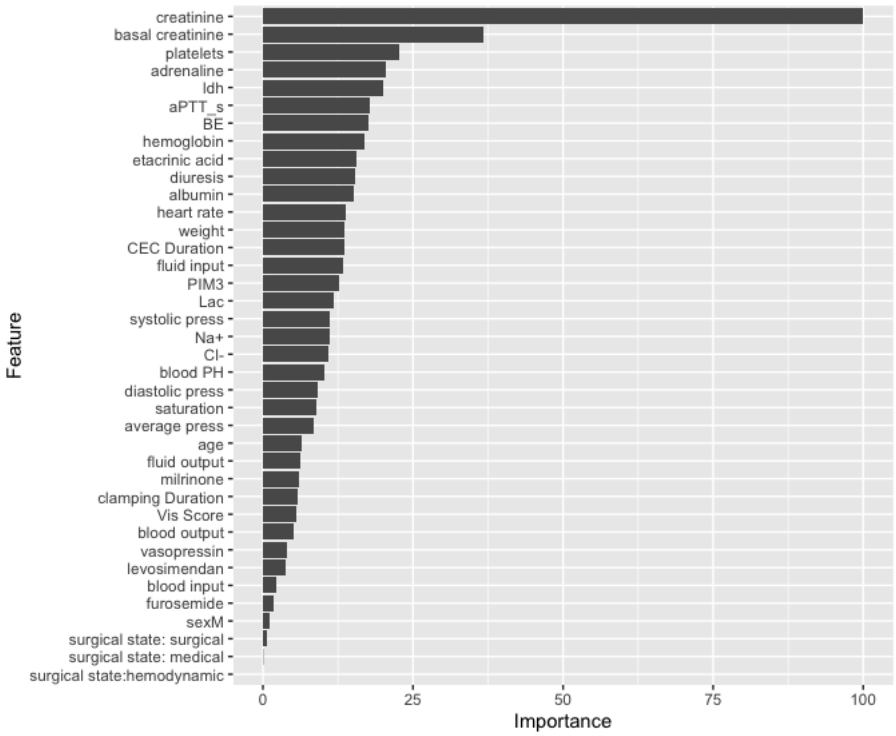

Figure 3.6: Variable importance case mode AKI

In both cases, although the result can be considered good, we observe the tendency to confuse "class 3", corresponding to the most severe stage 3 of AKI, with "class 0". This is a non-optimal result from a medical point of view.

3.5 Classification results using a subset of variables selected using RFE

Generally speaking, a simpler model is preferable to respect a complex one if it provides equivalent performance. From a statistical point of view, it is desirable to reduce the number of input variables in order to obtain a simpler model, reduce the computational cost of modeling and, in some cases, improve the performance of the model.

In this paragraph, we present the results obtained using the RFE algorithm.

As described in Chen et al. [2020] and Kuhn and Johnson [2019], variable selection becomes prominent, especially in data sets with many variables and features. It will eliminate unimportant variables and improve the accuracy as well as the performance of classification.

RFE is an algorithm for selecting features that are most relevant in predicting the target variable in a predictive model (either regression or classification). RFE applies a backward selection process to find the optimal combination of features.

First, it builds a model based on all features and calculates the importance of each feature in the model. Then, it rank-orders the features and removes the one(s) with the least importance iteratively based on model evaluation metrics (e.g. accuracy).

In figure 3.7 is shown the first version of this algorithm (Algorithm 1).

To get performance estimates that incorporate the variation due to feature selection, it is sug-

Algorithm 1: Recursive feature elimination

1.1 Tune/train the model on the training set using all predictors
1.2 Calculate model performance
1.3 Calculate variable importance or rankings
1.4 **for** *Each subset size* S_i, $i = 1 \ldots S$ **do**
1.5 Keep the S_i most important variables
1.6 [Optional] Pre–process the data
1.7 Tune/train the model on the training set using S_i predictors
1.8 Calculate model performance
1.9 [Optional] Recalculate the rankings for each predictor
1.10 **end**
1.11 Calculate the performance profile over the S_i
1.12 Determine the appropriate number of predictors
1.13 Use the model corresponding to the optimal S_i

Figure 3.7: RFE algorithm 1

gested that the steps in Algorithm 1 be encapsulated inside an outer layer of resampling (e.g. 10-fold cross-validation). Algorithm 2 shows a version of the algorithm that uses resampling.

Algorithm 2: Recursive feature elimination incorporating resampling

```
for Each Resampling Iteration do
    Partition data into training and test/hold–back set via resampling
    Tune/train the model on the training set using all predictors
    Predict the held–back samples
    Calculate variable importance or rankings
    for Each subset size S_i, i = 1...S do
        Keep the S_i most important variables
        [Optional] Pre–process the data
        Tune/train the model on the training set using S_i predictors
        Predict the held–back samples
        [Optional] Recalculate the rankings for each predictor
    end
end
Calculate the performance profile over the S_i using the held–back samples
Determine the appropriate number of predictors
Estimate the final list of predictors to keep in the final model
Fit the final model based on the optimal S_i using the original training set
```

Figure 3.8: RFE algorithm 2

The results shown below are computed in the following way:

- we use RFE algorithm 2 to select a subset of variables;
- we perform a classification procedure with a random forest algorithm implemented with the subset of variables identified previously.

Also in this case the performance is calculated on the test dataset not involved in the training phase. A comparison between the results obtained using RFE and all the variables will be discussed in the last paragraph of this chapter.

3.5.1 Binary AKI case using RFE

In this section, we show the results obtained by applying RFE in the binary AKI case. In figure 3.7 we show the corresponding values of accuracy obtained using the RFE method. As can be seen from the trend in the figure the selected number of variables is 15.

In Table 3.6 the list of the top 15 selected variables is inserted.

In Table 3.7 we insert the results obtained by applying the random forest classification algorithm to the selected subset of variables.

In figure 3.10 we show the obtained variable importance.

3.5.2 Severe AKI case using RFE

In this section, we show the results obtained in severe AKI cases.

In figure 3.11 we show the corresponding values of accuracy obtained using the RFE method. As can be seen from the trend in the figure the selected number of variables is 10.

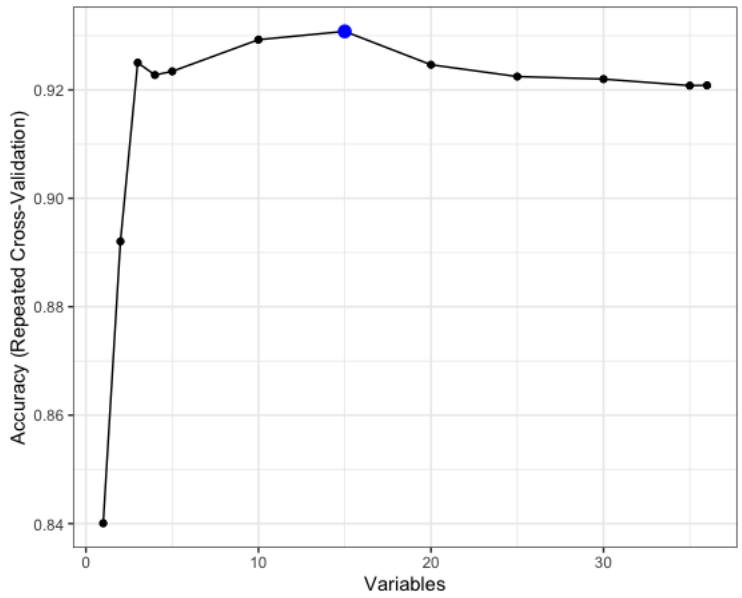

Figure 3.9: Accuracy plot binary case

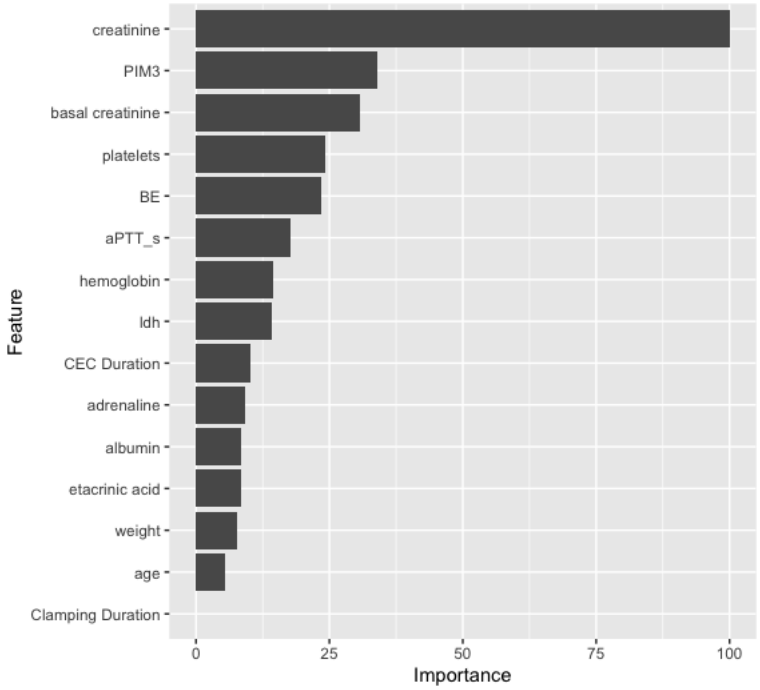

Figure 3.10: Variable importance binary AKI case

variables
creatinine
basal creatinine
ldh
etacrynic acid
platelets
aPTT_s
hemoglobin
CEC Duration
PIM3
adrenaline
albumin
weight
BE
age
Clamping Duration

Table 3.6: Top 15 variables: case binary AKI

	48h binary AKI score using RFE	CI
SENS	0.75	0.72-0.78
SPEC	0.98	0.97-0.99
TP	552	-
FP	61	-
TN	3059	-
FN	183	-
AUC-ROC	0.95	0.94-0.96

Table 3.7: 48h classification results via Random Forest using a subset of 15 variables selected using RFE (binary AKI case)

In Table 3.8 the list of the top 10 selected variables is inserted.

In Table 3.9 we insert the results obtained by applying the random forest classification algorithm to the selected subset of variables.

Lastly, in figure 3.12 we show the obtained variable importance.

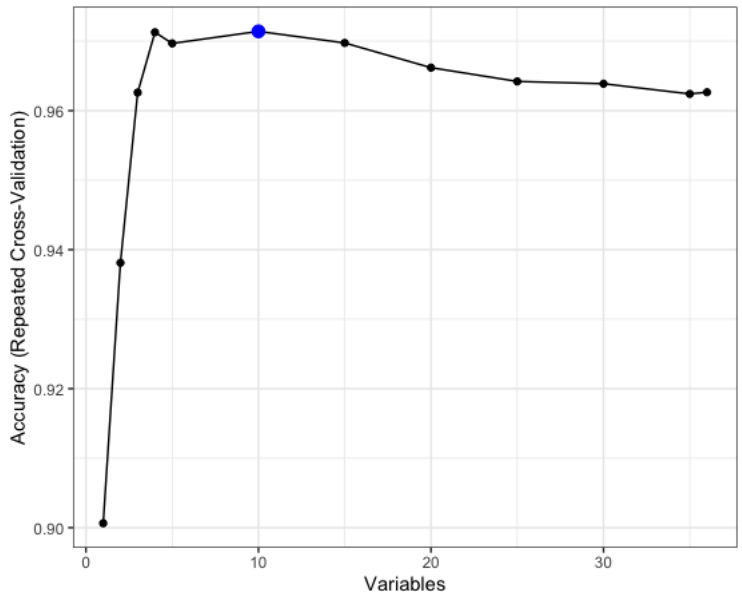

Figure 3.11: Accuracy plot severe AKI case

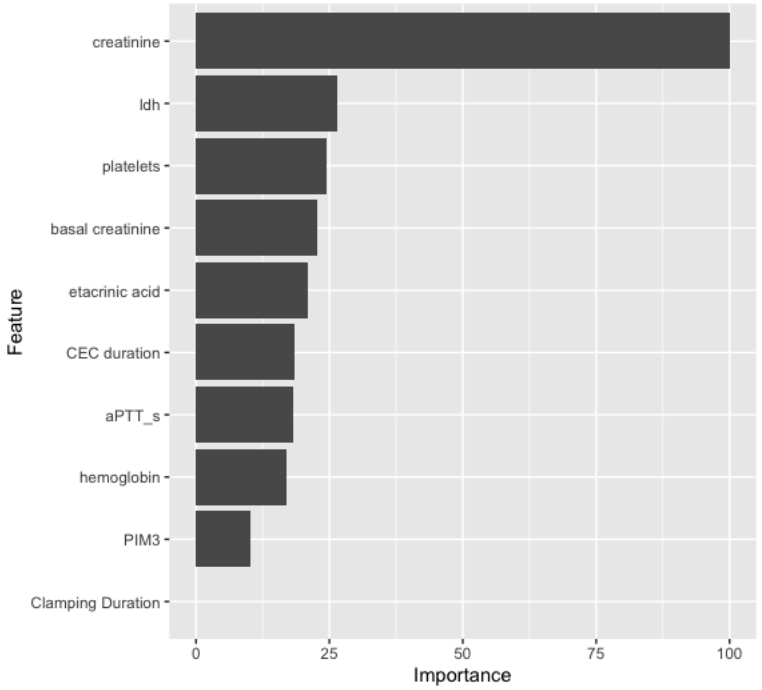

Figure 3.12: Variable importance severe AKI case using the subset of 10 variables selected with RFE.

variables
creatinine
basal creatinine
ldh
CEC Duration
platelets
etacrynic acid
aPTT_s
emoglobin
Clamping Duration
PIM3

Table 3.8: Top 10 variables: case severe AKI

	48h severe AKI score using RFE	CI
SENS	0.89	0.86-0.92
SPEC	0.99	0.99-1
TP	421	-
FP	35	-
TN	3347	-
FN	52	-
AUC-ROC	0.98	0.97-0.99

Table 3.9: 48h classification results via Random Forest using a subset of 10 variables selected via RFE (severe AKI case)

3.5.3 Max AKI case using RFE

In the paragraph, we show the results obtained in the max AKI case.

In figure 3.13 we show plots with the accuracy obtained using the RFE method. As can be seen from the trend in the figure the selected number of variables is 4.

In Table 3.12 the list of the top 4 selected variables is inserted.

variables
creatinine
basal creatinine
ldh
CEC Duration

Table 3.10: Top 4 variables: case max AKI

In Table 3.11 the confusion matrix in the maximum AKI case using the identified subset of variables is shown. The accuracy and the kappa for the max AKI case are the following:

- Accuracy: 0.93 (CI: 0.92, 0.93)
- Kappa: 0.76

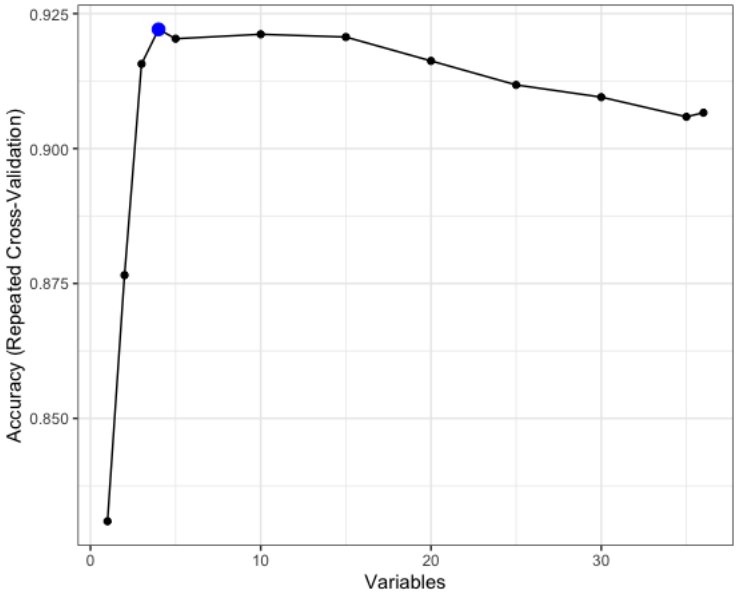

Figure 3.13: Accuracy plot max AKI case.

Max AKI confusion matrix				
	Observed			
Prediction	0	1	2	3
0	3043	121	29	31
1	50	134	6	5
2	12	6	148	7
3	15	0	3	243

Table 3.11: Confusion matrix: max AKI case using RFE

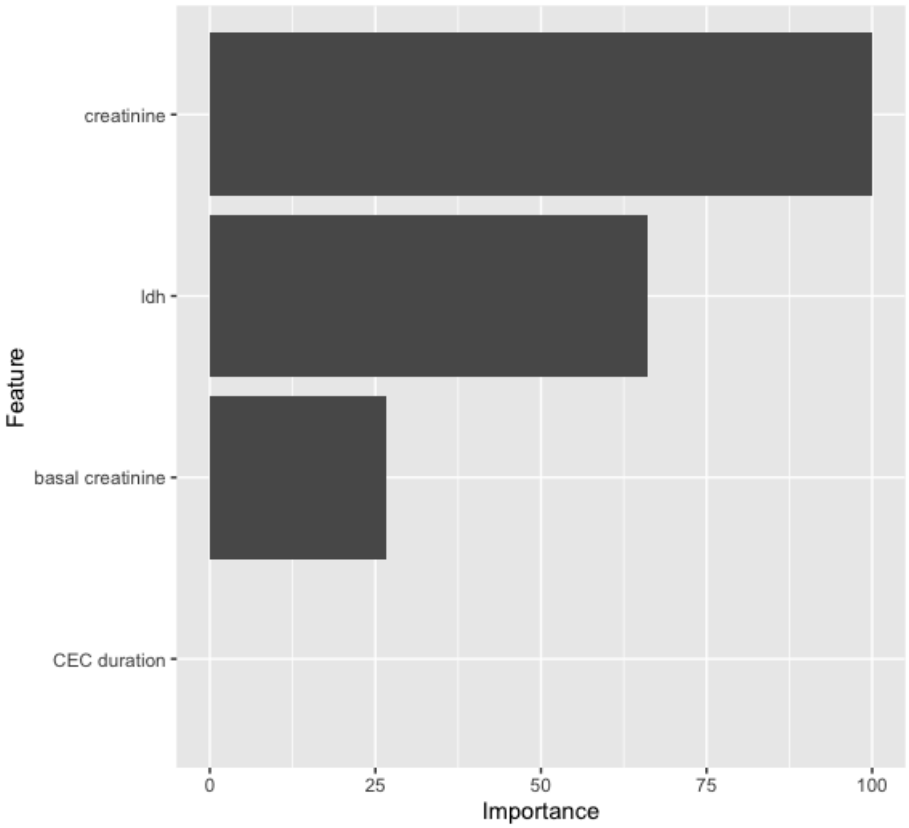

Figure 3.14: Variable importance max AKI case

In figure 3.14 we show the variable importance obtained in this case.
More detailed information is inserted in Appendix B.

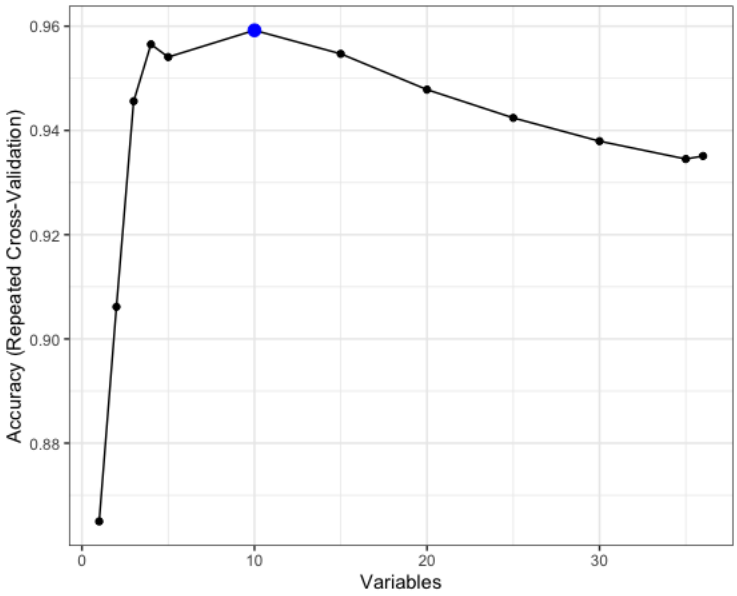

Figure 3.15: Accuracy plot mode AKI case

3.5.4 Mode AKI case using RFE

In this section, we show the results obtained in the mode AKI case.

In figure 3.15 we show a plot with the accuracy obtained using the RFE method. As can be seen from the trend in the figure the selected number of variables is 10.

In Table 3.12 the list of selected variables is inserted.

variables
creatinine
basal creatinine
ldh
CEC Duration
platelets
adrenaline
aPTT_s
hemoglobin
PIM3
albumin

Table 3.12: Top 10 variables: mode AKI case

In Table 3.13 the confusion matrix in the mode AKI case using the identified subset of variables is shown. The overall accuracy for the mode AKI case is the following:

- Accuracy: 0.96 (CI: 0.95, 0.97)
- Kappa: 0.85

In Figure 3.16 we insert the plot of variable importance for the mode AKI case.

More detailed information is inserted in Appendix B.

Mode AKI confusion matrix				
	Observed			
Prediction	0	1	2	3
0	3215	59	20	20
1	15	108	1	2
2	6	0	149	10
3	19	0	2	228

Table 3.13: Confusion matrix: mode AKI case using RFE

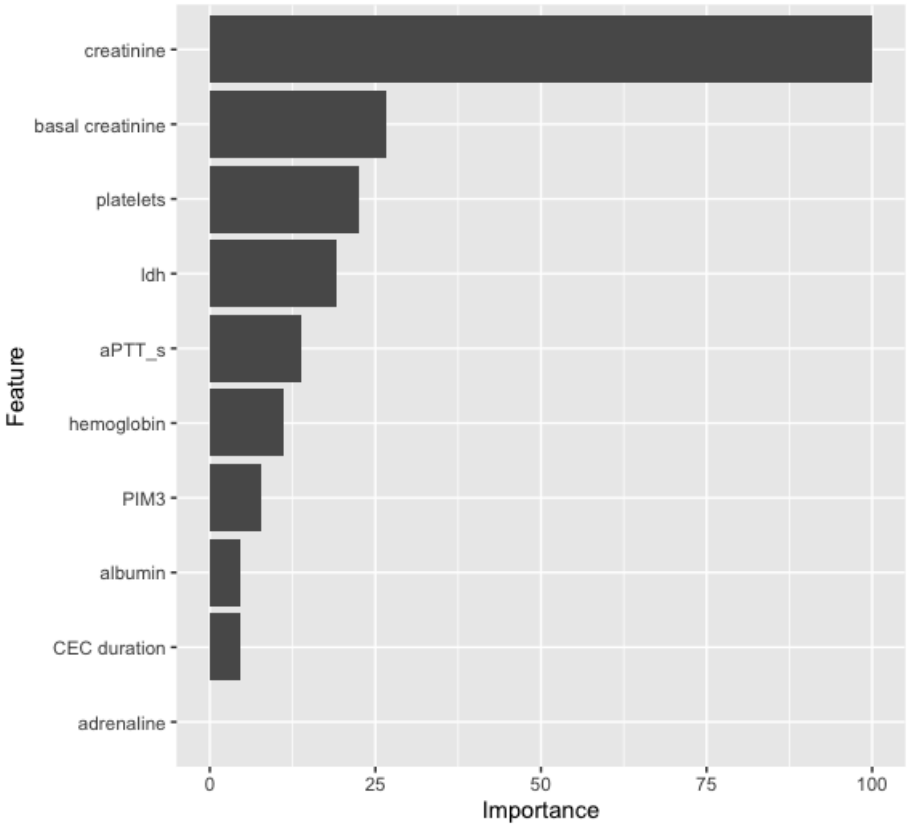

Figure 3.16: Variable importance mode AKI case using RFE

3.6 Analysis and comparison of results

In this section, we insert some final and comparative considerations on the results previously inserted.

First of all, we observe that both methods, the one that uses all the variables selected on the basis of the literature and the experience of doctors as potentially connected with the phenomenon of AKI and the type of pediatric patients of PCICU, and the one that uses a subset of variables, have good performance in terms of AUC-ROC and Accuracy.

In table 3.14 we show a summary of results with both approaches.

Comparison between results		
	Model with all variables	*Model with a subset of variables using RFE*
binary AKI	AUC-ROC 0.93 (0.92-0.94) SENS 0.71 (0.68-0.74) SPEC 0.99 (0.98-1)	AUC-ROC 0.95 (0.94-0.96) SENS 0.75 (0.72-0.78) SPEC 0.98 (0.97-0.99)
Severe AKI	AUC-ROC 0.99 (0.98-1) SENS 0.74 (0.70-0.78) SPEC 0.99 (0.99-1)	AUC-ROC 0.98 (0.97-0.99) SENS 0.89 (0.86-0.92) SPEC 0.99 (0.99-1)
Maximum AKI	Accuracy 0.92 (0.91, 0.93) Kappa 0.71	Accuracy 0.93 (0.92, 0.93) Kappa 0.76
Mode AKI	Accuracy 0.95 (0.94, 0.96) Kappa 0.80	Accuracy 0.96 (0.95, 0.97) Kappa 0.85

Table 3.14: Comparison between results obtained using the model with all variables and with a subset of variables selected using RFE algorithm.

Our opinion is that the results obtained using a subset of variables are slightly better and are preferred as they involve fewer variables and for this reason simpler interpretation by doctors. Moreover, in multiclass cases starting from the confusion matrix evaluated on the test dataset, we observe that results obtained after applying RFE tend to confuse less class 3 (the most severe stage of AKI) with class 0 (a patient without AKI). This aspect is also of potential interest to clinicians.

From the point of view of the importance of variables, we insert in the following figure a comparison between the first variables listed by importance in both approaches.

In figure 3.17 we insert a comparison between the first 10 variables in the case of binary AKI.

In figure 3.18 we show the case of severe AKI.

Instead, in the figures 3.20 and 3.19 we present the case respectively of max and mode AKI.

Although there are variations in the positions, we observe a substantial permanence in the first positions of some variables (for example creatinine, basal creatinine, ldh, etc.).

These encouraging results are believed to need further study and confirmation by applying them to a broader dataset of patients.

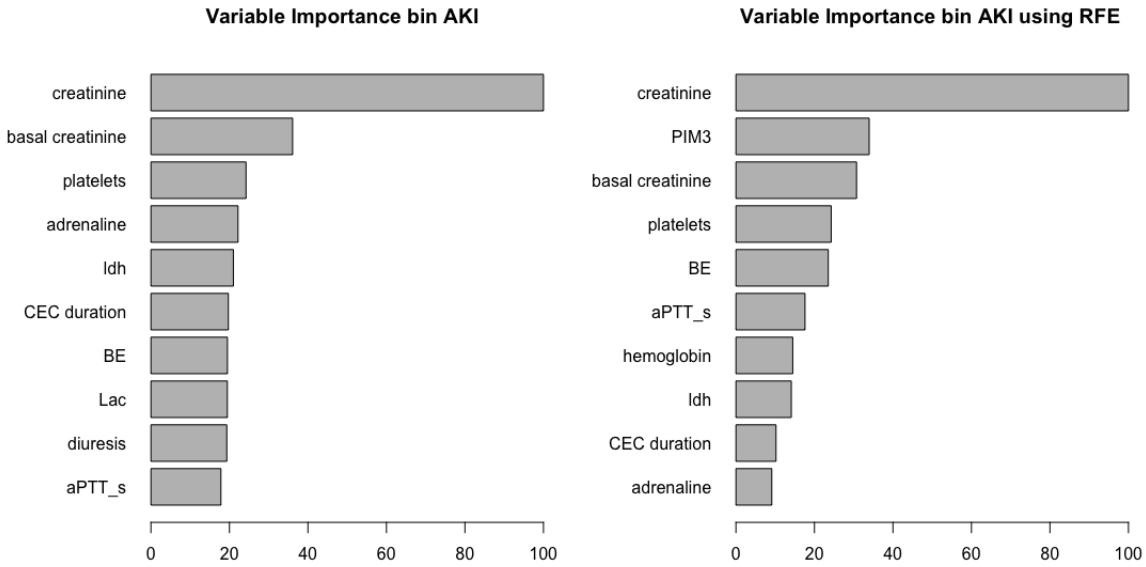

Figure 3.17: Comparison between the first 10 variables: binary AKI case.

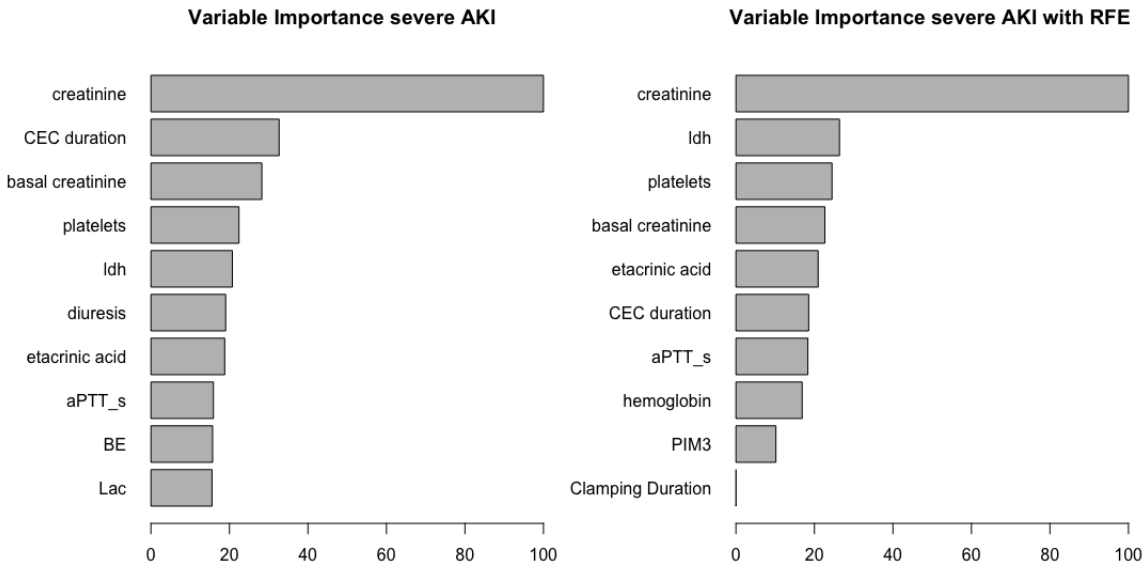

Figure 3.18: Comparison between the first 10 variables: severe AKI case.

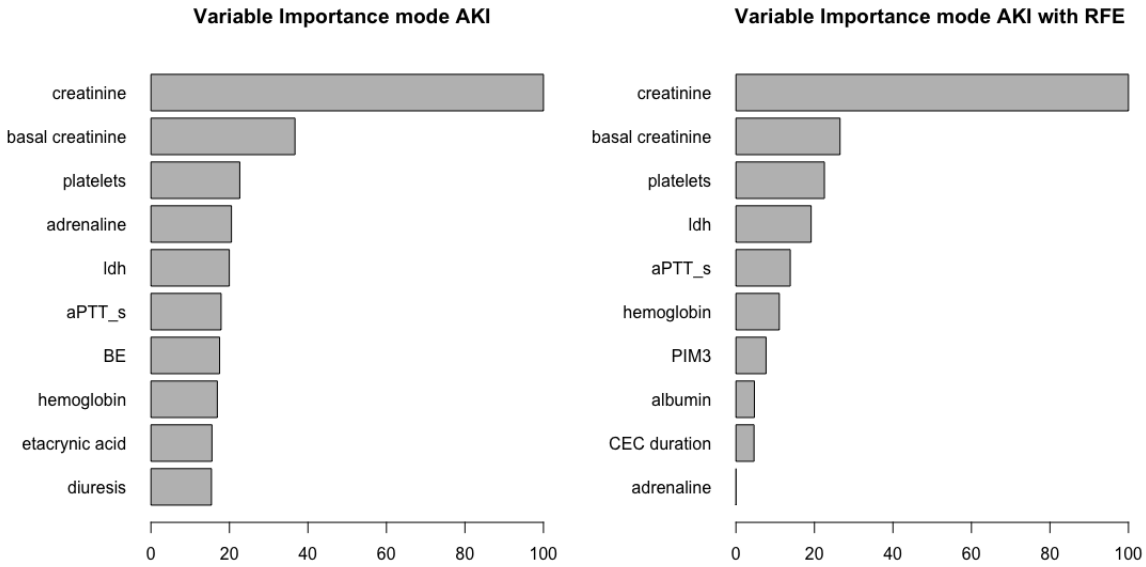

Figure 3.19: Comparison between the first 10 variables: mode AKI case.

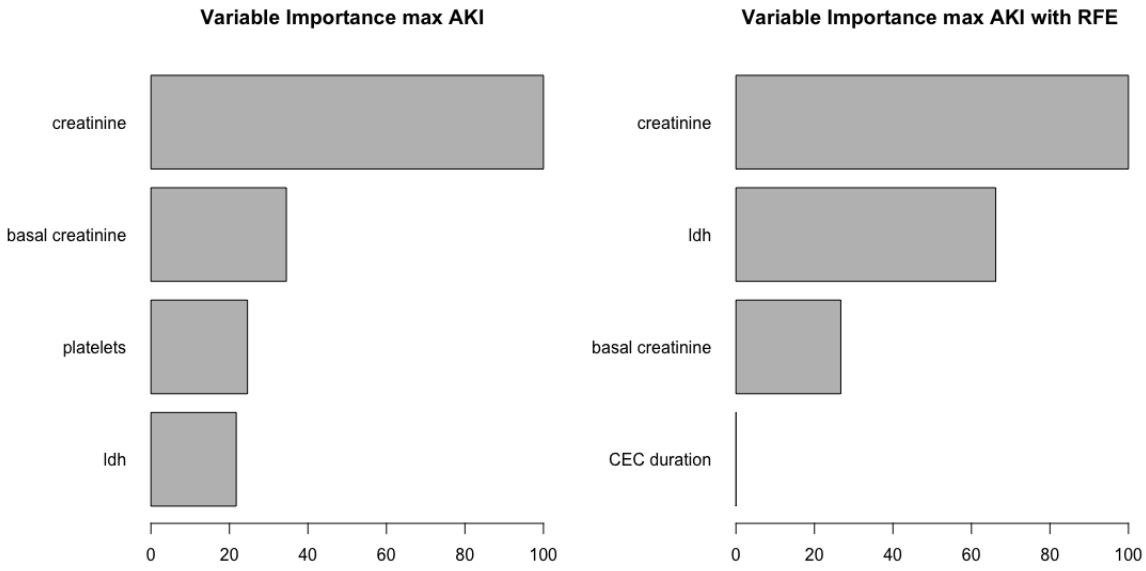

Figure 3.20: Comparison between the first 4 variables: max AKI case.

Chapter 4

Generalized Additive Models applied to PCICU

4.1 Introduction

After obtaining the results illustrated in the previous chapter, we shifted our reflection to trying to apply more interpretable models with the aim of eventually providing doctors additional information on the onset of AKI in addition to prediction and the variable importance. In fact, despite the results obtained using a Random Forest approach has shown very good results from a forecasting point of view, they lack interpretability. For this reason, we try to develop different approaches.

In this chapter we show some results obtained with the first chosen approach that is able to provide more interpretable results: Generalized Additive Models (GAMs).

A generalized additive model is a generalized linear model in which the response variable depends linearly on the sum of unknown smooth functions of some predictor variables, and interest focuses on inference about these smooth functions.

GAMs were originally developed by Hastie and Tibshirani [1986], to blend properties of generalized linear models (GLMs) with additive models.

The main idea is related to extending the linear regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon_i$$

enabling a non-linear relationship between each variable and the response.

In order to do that each linear component $\beta_j x_{ij}$ is replaced with a smooth non-linear function $f_j(x_{ij})$.

According to this, we can write the model as

$$y_i = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + ... + f_p(x_{ip}) + \epsilon_i$$

or more synthetically:

$$y_i = \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i.$$

The additive linear regression model is an example of a generalized additive model.

As described in Wood [2017], we can write, more generally, the GAM structure for example as:

$$g(\mu_i) = \mathbf{A}_i\boldsymbol{\theta} + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + ...$$

with $\mu_i = \mathbb{E}(Y_i)$ and $Y_i \sim EF(\mu_i, \Phi)$ where Y_i is the response variable, $EF(\mu_i, \Phi)$ an exponential family distribution with mean μ_i and scale parameter Φ, A_i is a row of the model matrix for any strictly parametric model components, $\boldsymbol{\theta}$ the corresponding parameter vector, and, finally, f_i are smooth functions of covariates x_k.

The use of GAMs over time has spread (Hastie [2017], Wood [2017]), they are widely used in practice (see Hastie and Tibshirani [1987], White et al. [2020], Ruppert et al. [2003]) and their popularity derives from the availability of smoothing parameter estimation methods statistically well-founded and numerically efficient (see Wood [2011], Wood et al. [2016]). So it is possible to estimate how smooth the component functions of a model should be.

As a means of doing this, an important parameter is the effective degrees of freedom (edf) estimated from generalized additive models; it is used as a proxy for the degree of non-linearity relationships.

We have that:

- An *edf* of 1 is equivalent to a linear relationship;
- an $1 < edf \leq 2$ is a weakly non-linear relationship;
- an $edf > 2$ indicates a highly non-linear relationship.

GAMs can also be used in contexts where y is qualitative. Assuming that y takes on values zero or one, and let $p(x) = Pr(y = 1|x)$ be the conditional probability (given the predictors) that the response equals 1, we can adapt the logistic regression model:

$$log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$$

extending this to take into account non-linear relationships in this natural way:

$$log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + f_1(x_1) + f_2(x_2) + ... + f_p(x_p)$$

A similar approach can be used in the multiclass case using multinomial distribution.

GAM is a powerful and flexible model setup since:

- it is interpretable;
- it allows the use of flexible predictor functions able to uncover hidden patterns in the data;
- it avoids overfitting via the regularization of predictor functions.

As reported in Hastie et al. [2001], GAMs has also a limitation, deriving from the characteristic of an additive model. With many variables, important interactions can be missed. However, as with linear regression, it is possible to add manually some interaction terms by including additional predictors of the form $X_j \times X_k$.

We fit a generalized additive model (GAM) to our dataset with `mgcv` R package using the same approach followed by the random forest classification.

In particular we:

- split the dataset in learning and test (70% for learning and 30% for test); in splitting the data, we preserve the percentages of each class in train and test sets;
- apply it to all the available variables;
- categorical variables (sex and surgical state) are coded as factors;
- employs GAM for both binary and multiclass classification cases;
- use GAM logistic regression for binary cases and GAM multinomial logistic regression for multiclass cases;
- selecting smoothing parameters by REML (less prone to local minima than the other criteria as described in Wood [2017]).

4.2 Results with binary classification

In this paragraph, we present the results in the binary classification case. We show the results using the approach followed in chapter 3.

- binary AKI case;
- severe AKI case.

We apply the GAM model using all the variables used in the case of random forest.

Further attempts made by eliminating non-significant variables or considering effects due to interactions with the variables (e.g interaction between creatinine and basal creatinine like s(creatinine, basal creatinine) +s(creatinine)+ s(basal creatinine)) did not improve the results of the classification. For this reason, they are not reported here.

4.2.1 Binary AKI case

In this paragraph, we show the results for the binary AKI case using all the variables.

Starting from Table 4.1 it is possible to notice that:

- with the exception of a few (surgical state, ldh, average press, blood input, blood output), most of the additive components are significant;
- for many additive components the corresponding edf shows a highly non-linear relationship.

The univariate functional form of each additive component is depicted in the Figures: 4.1, 4.2, 4.3, 4.4. In these plots, the vertical axis is converted to the probability scale of developing AKI.

Among the various trends, some lend themselves to interpretations expected or compatible with the physiology of the AKI.

In particular, we observe that an increase in creatinine is linked to an increased risk of developing AKI, meanwhile greater values of basal creatinine show a protective effect.

Also in the case of diuresis, the first part of the curve of this variable has a clear interpretation given that the AKI is basically a kidney problem.

A. parametric coefficients	**Estimate**	**Std. Error**	**t-value**	**p-value**
(Intercept)	-2.199	0.097	-22.733	< 0.001
sexM	0.111	0.087	1.241	0.215
surgical_Surgical	0.065	0.119	0.551	0.582
surgical_Hemodynamic	-0.170	0.188	-0.908	0.364
surgical_Medical	-0.284	0.197	-1.440	0.150
B. smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
s(weight)	6.032	7.115	30.139	< 0.001
s(age)	5.820	6.866	50.795	< 0.001
s(basal creatinine)	6.617	7.611	424.422	< 0.001
s(PIM3)	5.679	6.64	41.90	< 0.001
s(Vis Score)	3.919	4.799	11.858	0.033
s(CEC Duration)	7.607	8.454	53.783	< 0.001
s(Clamping Duration)	1.000	1.000	9.999	0.002
s(diuresis)	7.056	7.780	90.234	< 0.001
s(fluid input)	4.773	5.915	43.424	< 0.001
s(fluid output)	6.725	7.439	68.069	< 0.001
s(creatinine)	3.290	3.910	418.703	< 0.001
s(ldh)	2.171	2.685	5.002	0.127
s(albumin)	3.801	4.748	10.715	0.042
s(hemoglobin)	5.088	6.181	16.843	0.011
s(platelets)	4.427	5.471	23.090	< 0.001
s(aPTTs)	4.808	5.873	33.726	< 0.001
s(BE)	2.330	3.031	8.159	0.044
s(Cl)	5.978	7.180	28.598	0.002
s(Lac)	1.873	2.371	8.707	0.024
s(Na)	3.560	4.559	6.441	0.234
s(blood PH)	2.250	2.914	6.953	0.089
s(blood input)	1.001	1.003	0.833	0.362
s(blood output)	2.808	3.476	3.523	0.313
s(systolic press)	6.449	7.634	29.593	0.002
s(diastolic press)	3.651	4.642	10.059	0.053
s(average press)	1.000	1.000	0.759	0.384
s(heart rate)	3.959	4.989	31.373	< 0.001
s(saturation)	3.597	4.513	14.533	0.013
s(adrenaline)	7.286	8.031	43.899	< 0.001
s(furosemide)	3.301	4.106	16.820	0.002
s(ethacrynic acid)	1.001	1.001	23.390	< 0.001
s(milrinone)	2.485	2.872	16.557	0.001
s(levosimendan)	1.129	1.241	10.881	0.003
s(vasopressin)	1.001	1.002	4.622	0.032

Table 4.1: Summary of GAM model (binary AKI case) with all variables

The classification results, in this case, are shown in Table 4.2 considering the reference class AKI 1. We obtain:

- AUC-ROC = 0.87
- sensitivity =0.51
- specificity= 0.97

More detailed information on performance is reported in Appendix C.

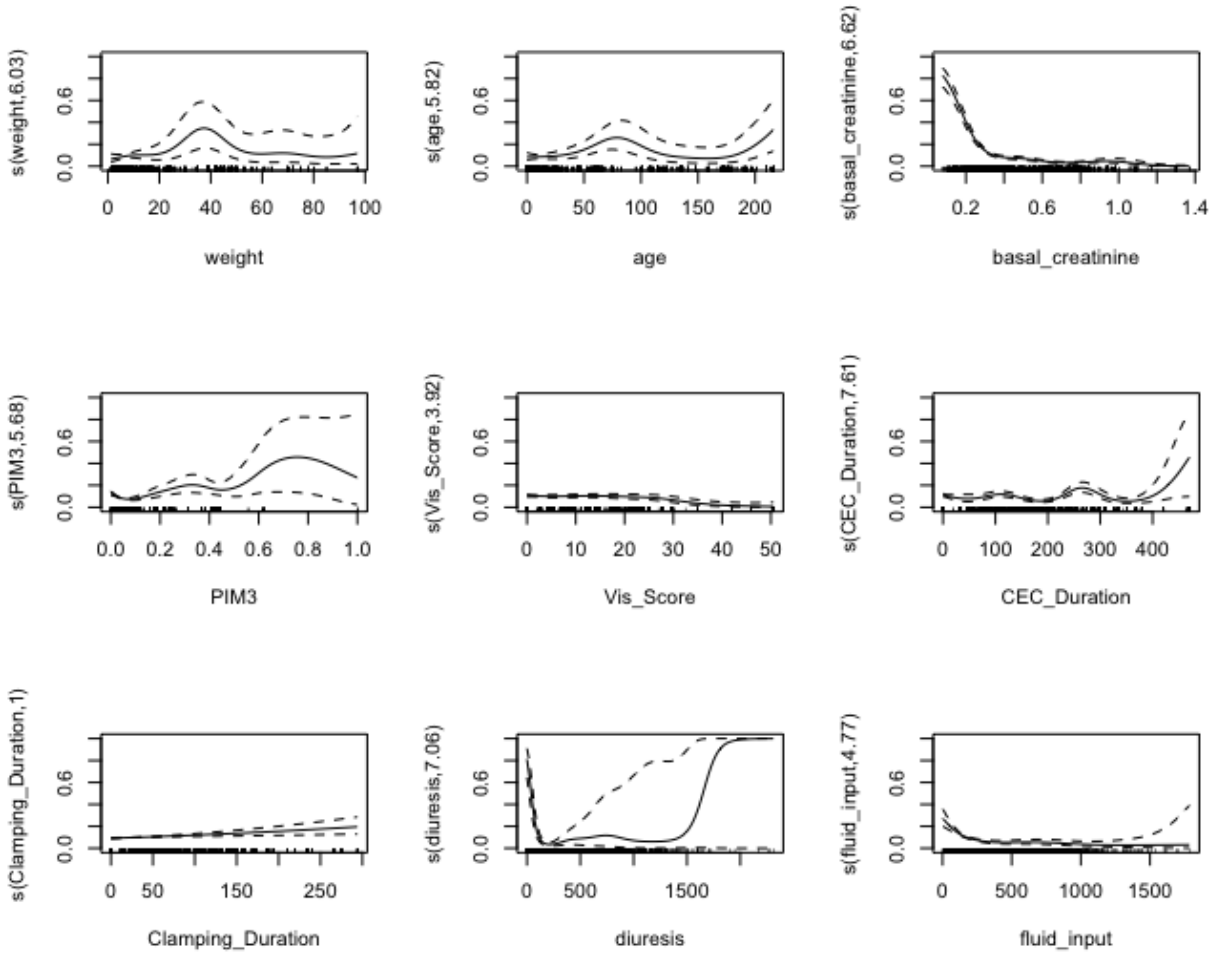

Figure 4.1: GAM bin AKI (plot1)

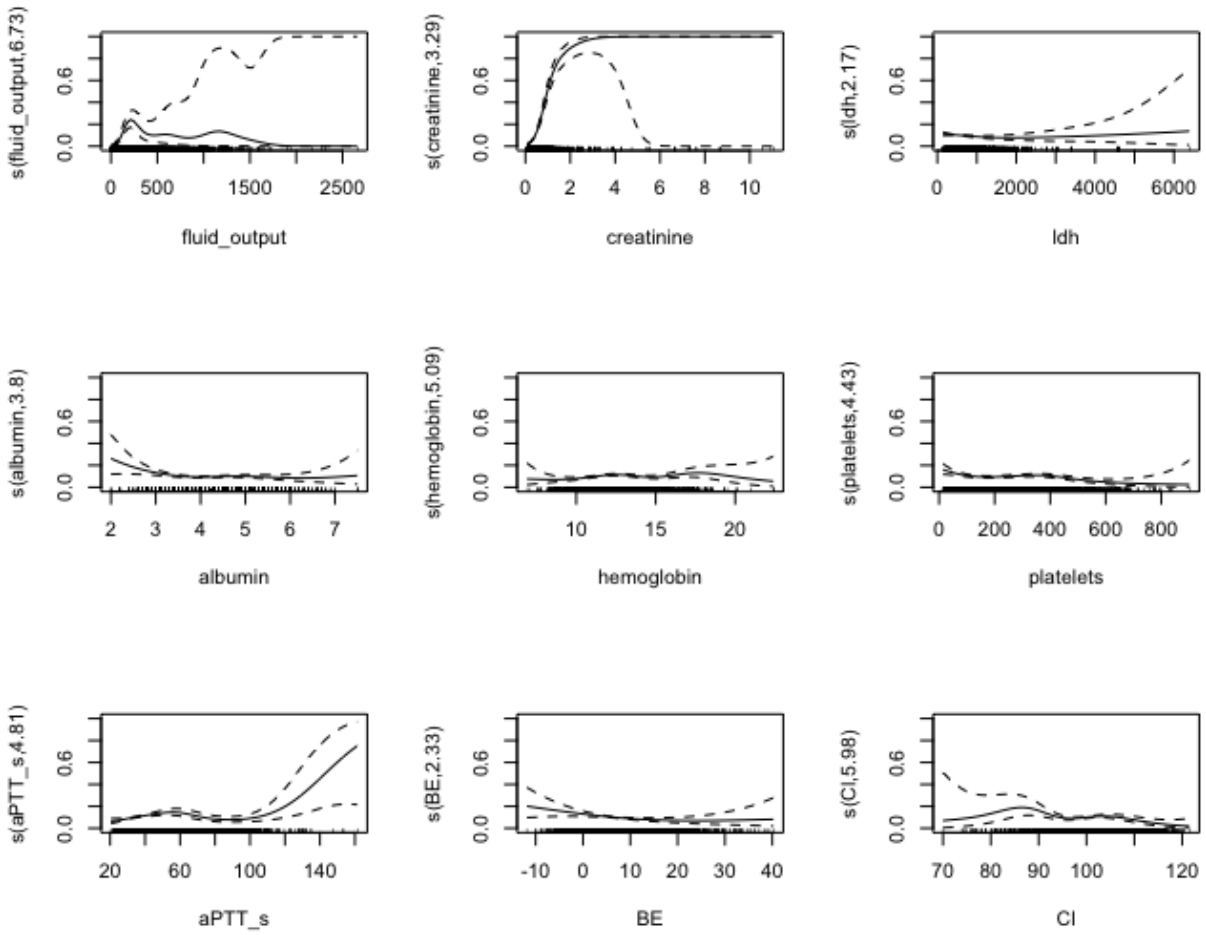

Figure 4.2: GAM bin AKI (plot2)

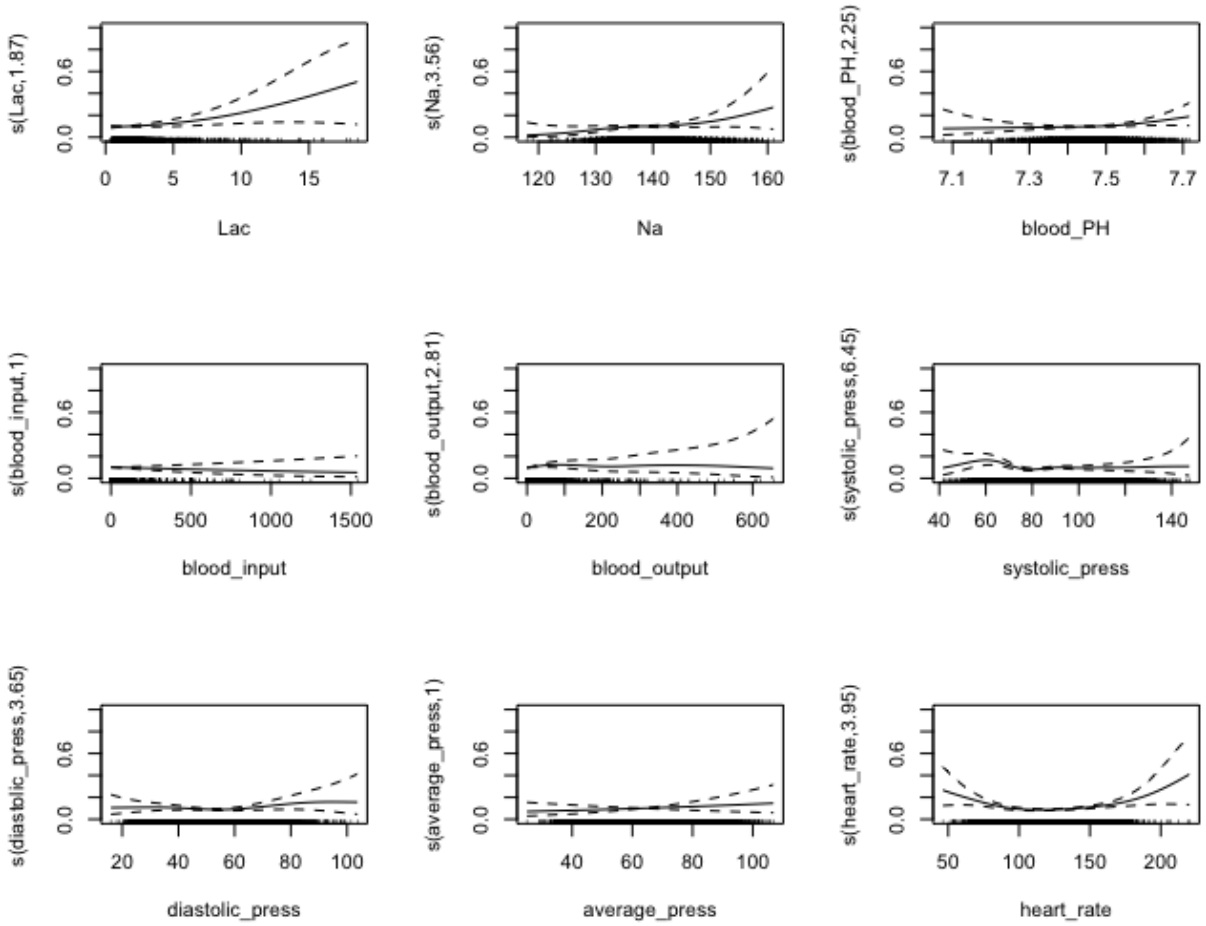

Figure 4.3: GAM bin AKI (plot3)

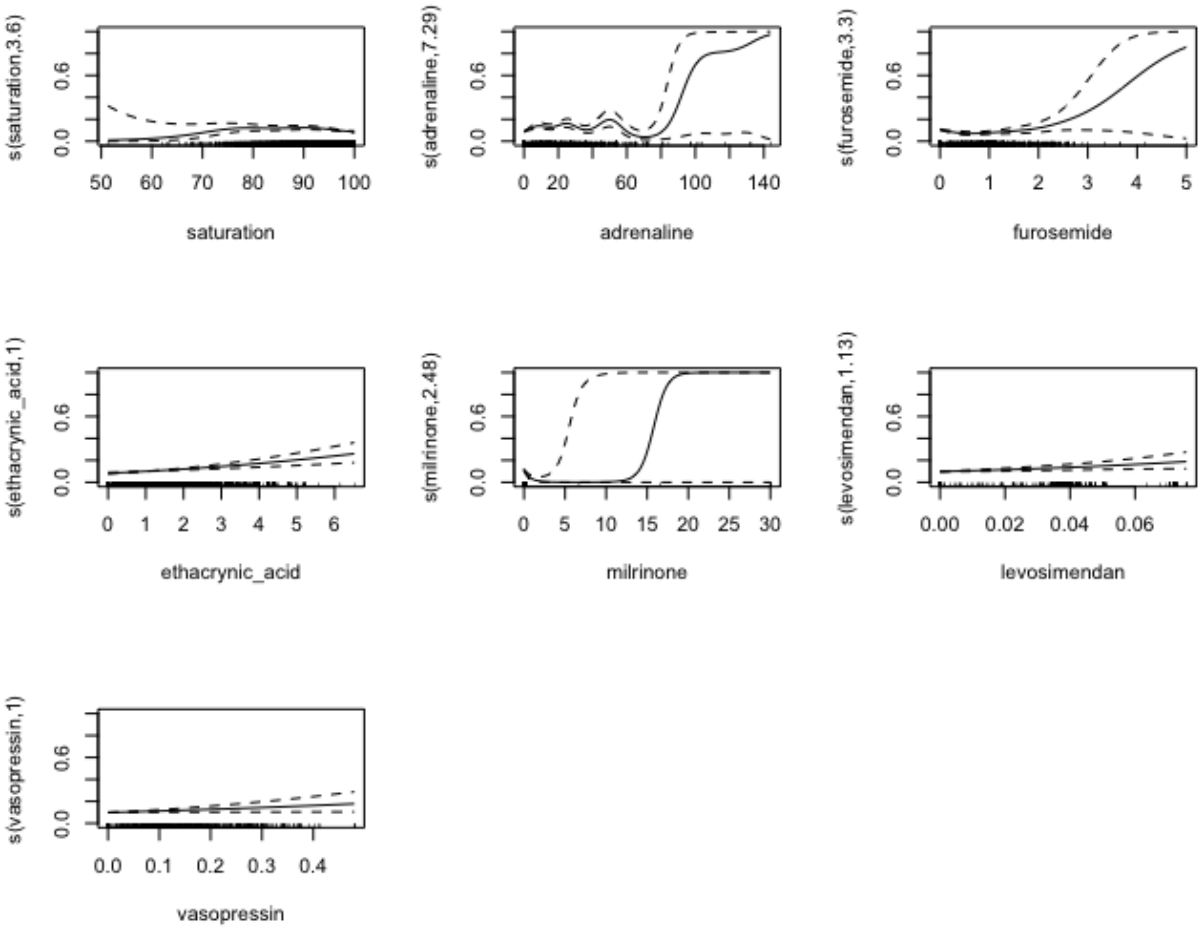

Figure 4.4: GAM bin AKI (plot4)

	48h binary AKI score using GAM	**CI**
SENS	0.51	0.47-0.55
SPEC	0.97	0.96-0.98
TP	375	-
FP	109	-
TN	3011	-
FN	360	-
AUC-ROC	0.87	0.85-0.89

Table 4.2: AUC-ROC of GAM model (binary AKI case with all variables)

4.2.2 Severe AKI case

In this paragraph, we show the results for predicting severe AKI case using all the variables. We notice that the observations that can be made are similar to the case of binary AKI.

Indeed, starting from Table 4.3 it is possible notice that:

- with the exception of a few (ldh, average press,Na, blood input, blood output), most of the variables are significant;
- many additive components show a highly non-linear relationship ($edf > 2$).

The trend of each variable is depicted in the following Figures: 4.5, 4.6, 4.7, 4.8.

In the case of non-linear variables, a trend similar to that obtained in the case of binary AKI is confirmed as well as the possible interpretation compatible with the physiology of the AKI.

Also in this case we observe that an increase in creatinine is linked to an increased risk of developing AKI; meanwhile, greater values of basal creatinine show a protective effect.

Also in the case of diuresis, the first part of the curve of this variable has a clear interpretation given that the AKI is basically a kidney problem.

A. parametric coefficients	**Estimate**	**Std. Error**	**t-value**	**p-value**
(Intercept)	-4.148	0.250	-16.575	< 0.001
sexM	0.441	0.138	3.188	< 0.001
surgical_state_Surgical	-0.398	0.174	-2.289	0.022
surgical_state_Hemodynamic	-0.860	0.285	-3.018	0.003
surgical_state_Medical	-0.435	0.299	-1.456	0.145
B. smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
s(weight)	6.365	7.462	31.101	0.001
s(age)	5.605	6.647	19.288	0.006
s(basal_creatinine)	7.596	8.374	399.493	< 0.001
s(PIM3)	4.525	5.447	41.695	< 0.001
s(Vis_Score)	6.399	6.932	23.263	< 0.001
s(CEC_Duration)	7.530	8.378	49.890	< 0.001
s(Clamping_Duration)	1.440	1.744	0.444	0.68
s(diuresis)	7.124	7.768	124.803	< 0.001
s(fluid_input)	5.367	6.587	28.519	< 0.001
s(fluid_output)	6.566	7.169	73.643	< 0.001
s(creatinine)	4.118	4.534	483.669	< 0.001
s(ldh)	2.902	3.568	8.198	0.049
s(albumin)	6.627	7.574	47.637	< 0.001
s(hemoglobin)	6.541	7.607	35.783	< 0.001
s(platelets)	4.365	5.400	15.544	0.012
s(aPTT_s)	4.367	5.361	46.120	< 0.001
s(BE)	3.685	4.720	7.838	0.148
s(Cl)	5.733	6.902	21.979	0.003
s(Lac)	1.020	1.039	16.286	<0.001
s(Na)	1.000	1.000	0.391	0.532
s(blood_PH)	3.739	4.743	24.856	0.001
s(blood_input)	1.000	1.001	0.287	0.592
s(blood_output)	1.533	1.887	1.124	0.599
s(systolic_press)	1.000	1.000	0.002	0.989
s(diastolic_press)	1.001	1.003	3.475	0.063
s(average_press)	3.213	4.097	12.650	0.014
s(heart_rate)	4.060	5.110	21.256	0.001
s(saturation)	1.000	1.000	4.871	0.027
s(adrenaline)	6.459	7.338	29.622	< 0.001
s(furosemide)	2.157	2.678	3.959	0.264
s(ethacrynic_acid)	1.000	1.000	10.619	0.001
s(milrinone)	2.078	2.562	0.931	0.832
s(levosimendan)	1.000	1.000	17.217	< 0.001
s(vasopressin)	1.000	1.000	3.949	0.047

Table 4.3: Summary of GAM model severe AKI case with all variables

The classification results, in this case, are shown in Table 4.4 considering the reference class AKI 1. We obtain:

- AUC-ROC = 0.94
- sensitivity =0.57
- specificity= 0.98

The value of the AUC-ROC is better than the value in the case of binary AKI although lower than that obtained with the Random Forest classification approach (see 3.6). More detailed information on the classification performance is reported in Appendix C.

	48h severe AKI score using GAM	**CI**
SENS	0.57	0.53-0.62
SPEC	0.98	0.97-0.99
TP	271	-
FP	71	-
TN	3311	-
FN	202	-
AUC-ROC	0.94	0.92-0.96

Table 4.4: AUC-ROC of GAM model (severe AKI case with all variables)

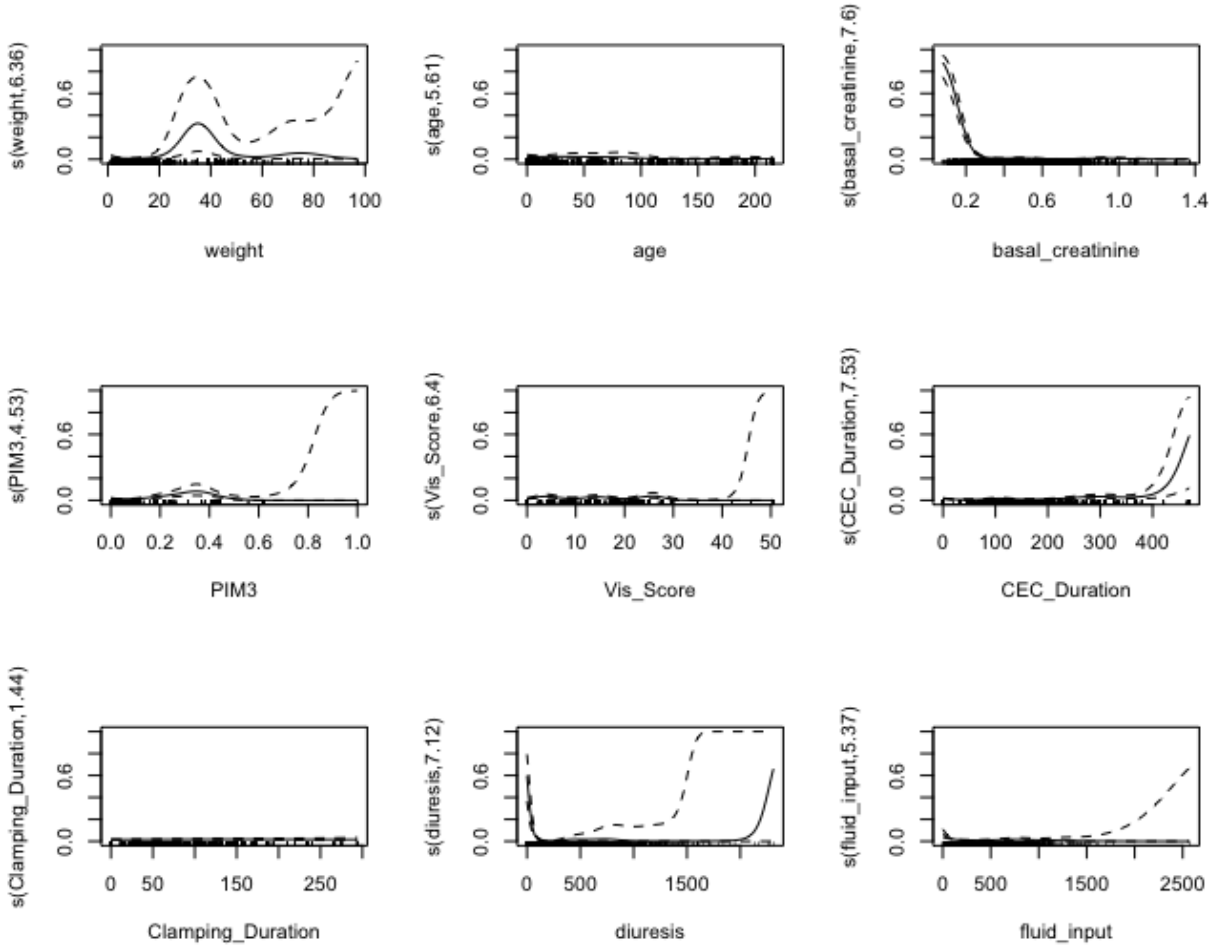

Figure 4.5: GAM severe AKI (plot1)

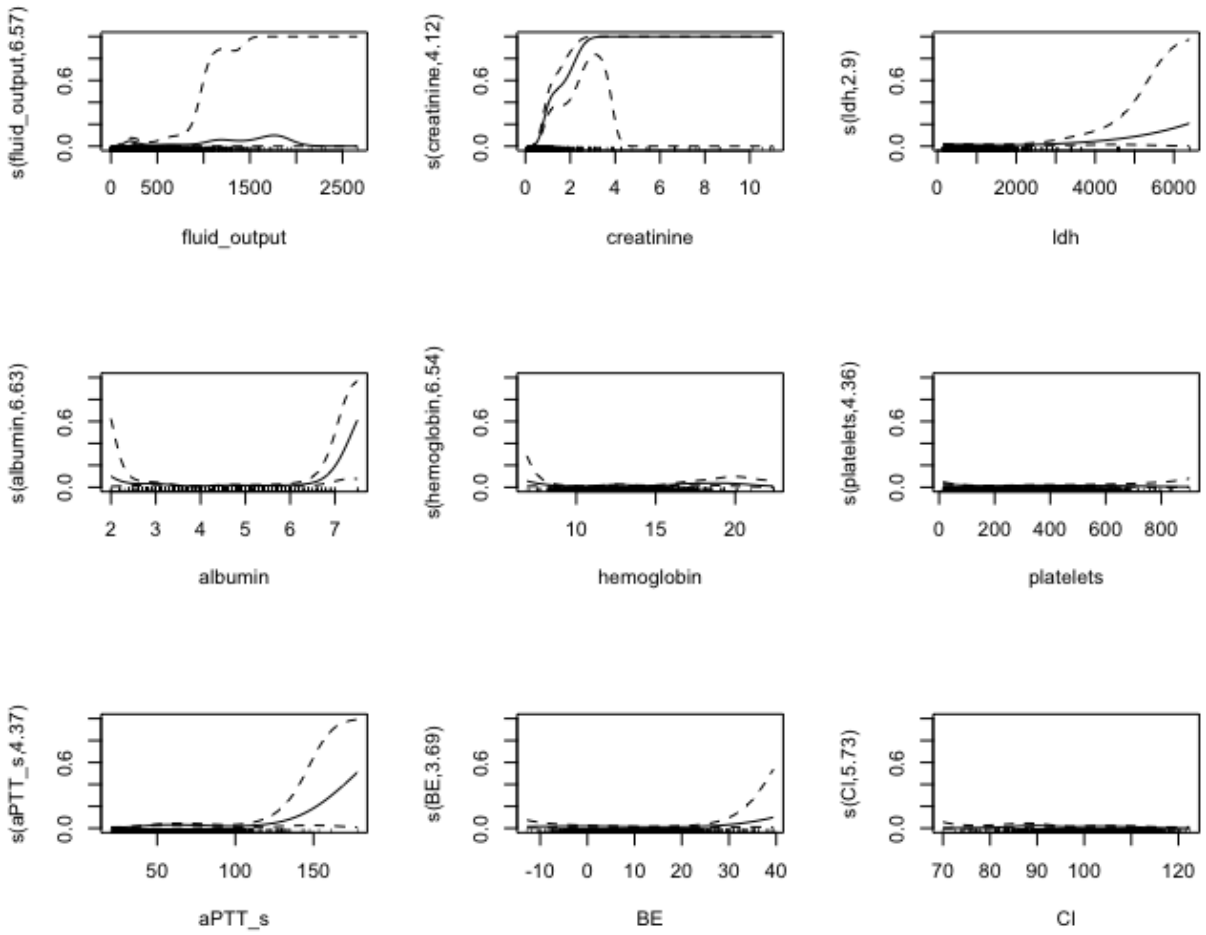

Figure 4.6: GAM severe AKI (plot2)

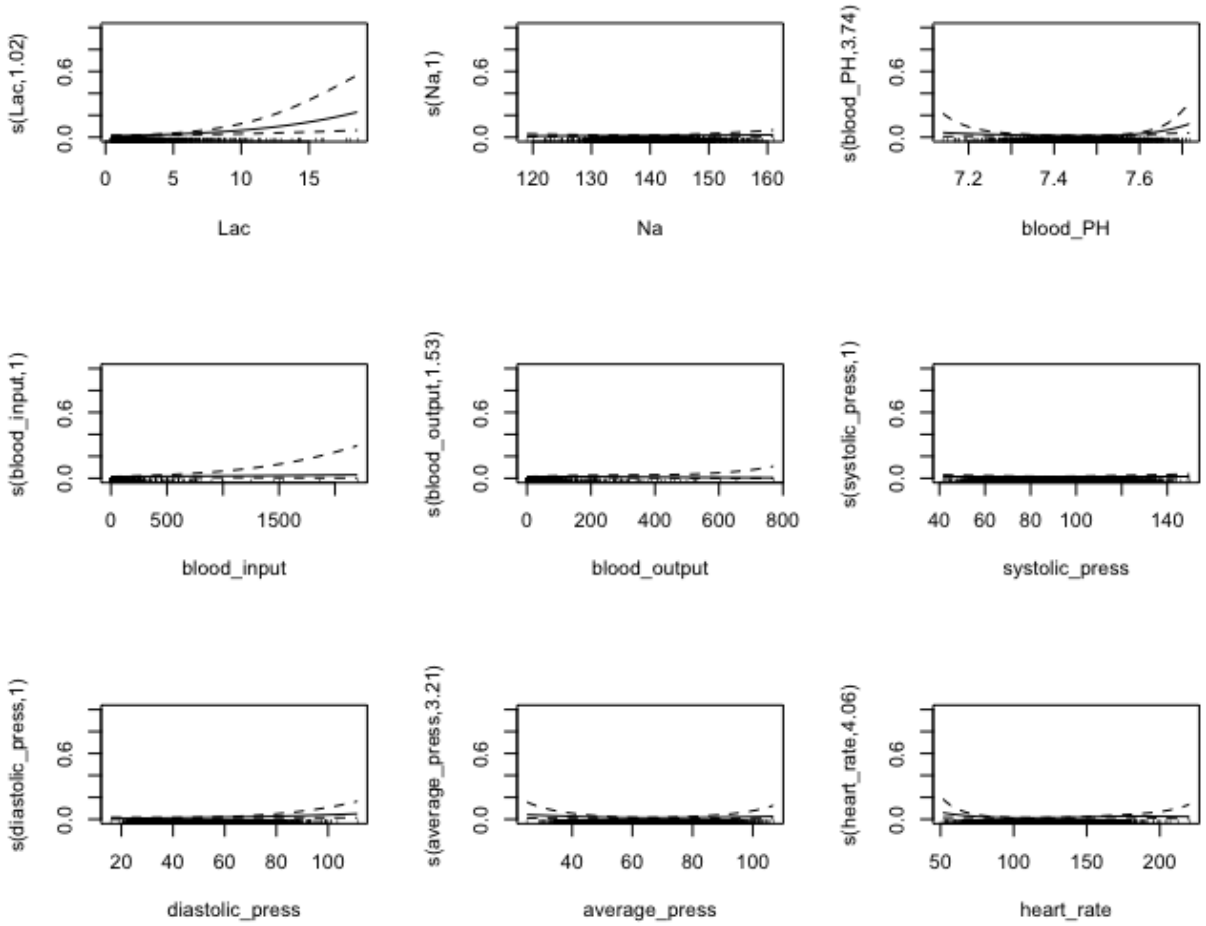

Figure 4.7: GAM severe AKI (plot3)

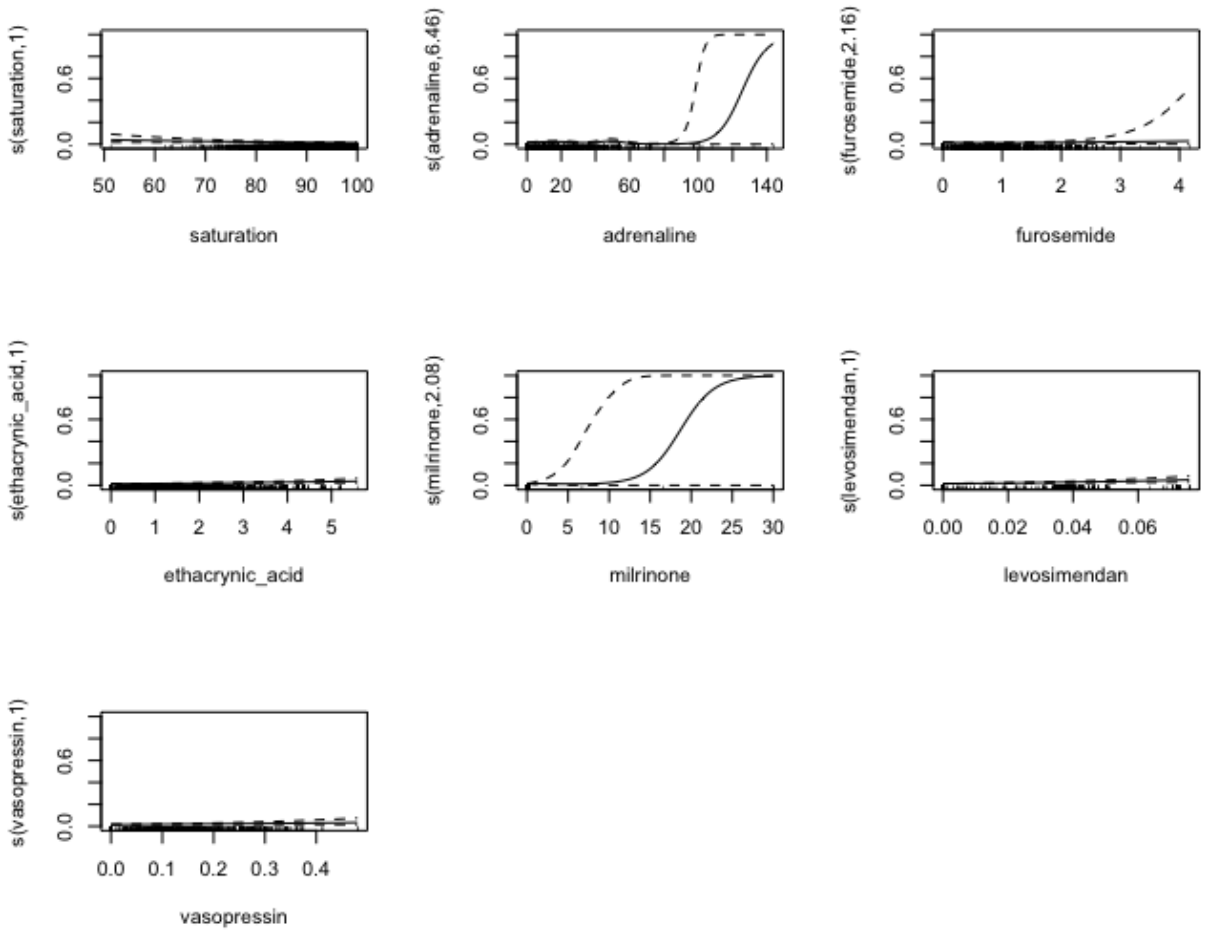

Figure 4.8: GAM severe AKI (plot4)

4.3 Multiclass case: mode and maximum AKI case

In this paragraph, we show the results obtained in the multiclass case considering the mode and maximum AKI case like in 3.4.2.

In this case, we use GAM multinomial logistic regression with class 0 as reference. For this reason, visualizing the model's results we obtain three plots for each variable of the model.

For the sake of brevity, we insert in Figure 4.9 an example of plots involving the variables creatinine, basal creatinine and diuresis considering mode AKI case.

It is interesting to note the variation in the trend of each variable as the AKI class increases (from the least severe to the most severe).

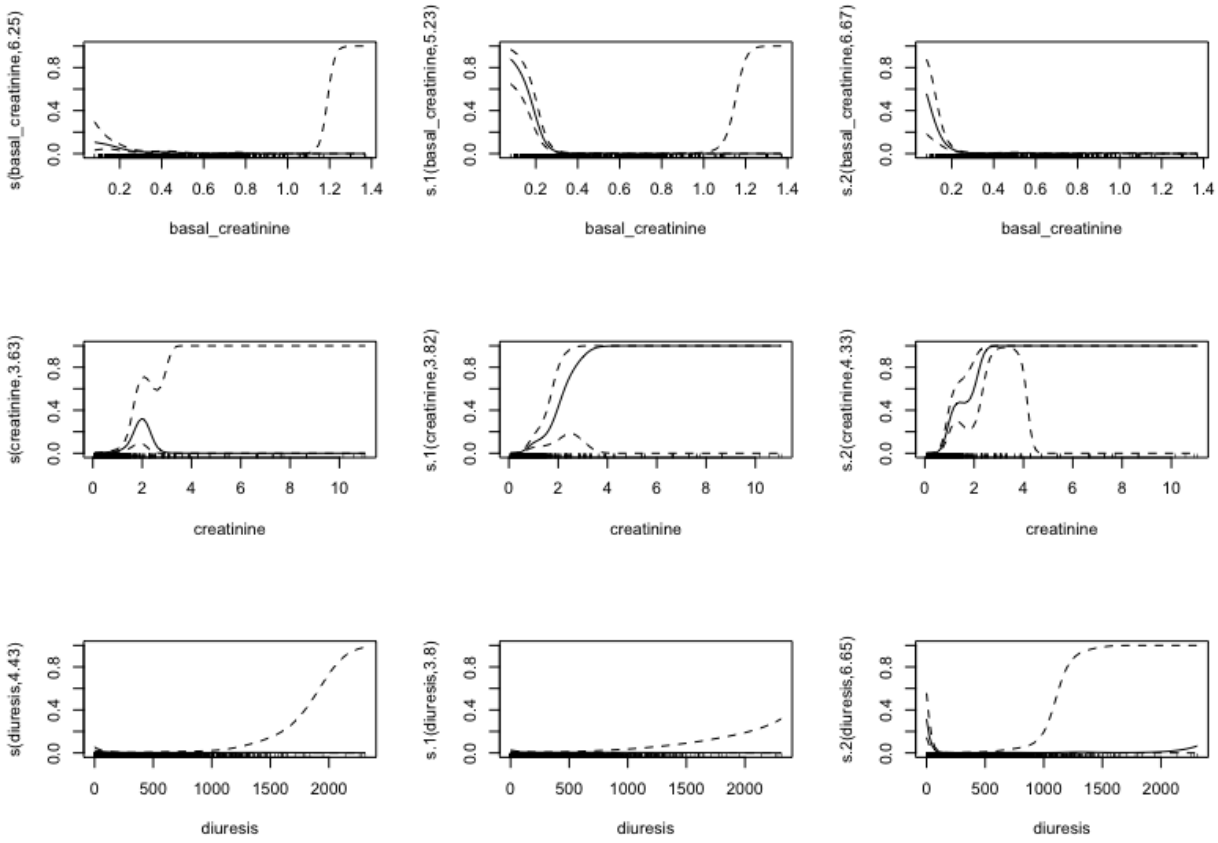

Figure 4.9: GAM mode AKI plot involving creatinine, basal creatinine, and diuresis

We report the confusion matrix achieved and the corresponding values of accuracy and kappa. In Table 4.5 is reported the confusion matrix obtained in the mode AKI case.

In particular, we have:

- Accuracy= 0.90 C.I. (0.89, 0.91)
- Kappa = 0.59

Mode AKI confusion matrix				
	Observed			
Prediction	0	1	2	3
0	3189	120	77	74
1	12	22	1	4
2	27	10	86	6
3	30	4	11	182

Table 4.5: Confusion matrix: mode AKI case using GAM approach

From the confusion matrix, we can see that the class 1 (the least worrying from a medical point of view) is often confused with class 0 (absence of AKI).

In Figure 4.10 we show an example of plot involving the variables creatinine, basal creatinine, and diuresis considering max AKI case.

Also in this case it is interesting to note the variation in the trend of each variable as the AKI class increases (from the least severe to the most severe).

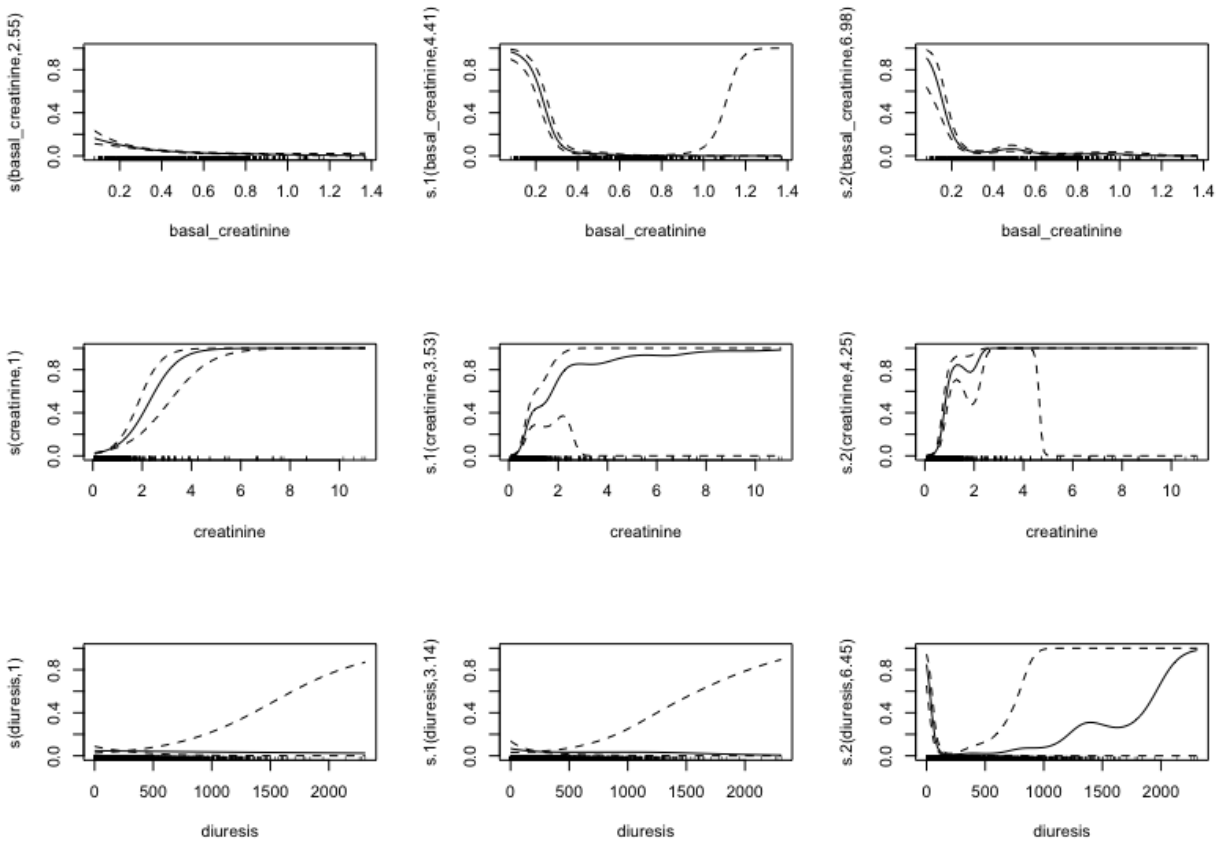

Figure 4.10: GAM max AKI plot involving creatinine, basal creatinine, and diuresis

In Table 4.6 we report the confusion matrix obtained in the max AKI case.
Concerning Accuracy and Kappa, we have:

- Accuracy= 0.87 C.I. (0.86, 0.89)
- Kappa = 0.54

The results are worse than the previous case and confirm the classification problems of AKI classes 1 and 2.

Max AKI confusion matrix				
	Observed			
Prediction	0	1	2	3
0	3056	222	82	80
1	14	19	1	2
2	22	11	93	9
3	30	2	9	203

Table 4.6: Confusion matrix: max AKI case using GAM approach

As shown in the previous tables, the results are not good enough in discriminating between classes 1 and 2 for both max and mode AKI cases.

For this reason, the results are reported only in a synthetic way in this paragraph.

4.4 Comparison between most important variables using Random Forests and GAMs

In this section, we compare the most important variables obtained with the RF classification method and the p-value of the corresponding variables in the GAMs.

The aim is to verify if there is a correspondence between the most important variables highlighted with the Random Forest approach and the same variables in the GAMs.

We apply this comparison only in the binary class case.

Table 4.7 is related to the binary AKI case.

We insert:

- the first ten most important variables for both cases: RF with all the variables and RF with a subset of variables selected using RFE method as reported in section 3.6;
- the corresponding p-value obtained with GAMs.

RF with all variables	**GAM (p-value)**	**RF with subset via RFE**	**GAM (p-value)**
creatinine	$p < 0.001$	creatinine	$p < 0.001$
basal creatinine	$p < 0.001$	PIM3	$p < 0.001$
platelets	$p < 0.05$	basal creatinine	$p < 0.001$
adrenaline	$p < 0.001$	platelets	$p < 0.05$
ldh	$p = 0.127$	BE	$p < 0.05$
CEC duration	$p < 0.001$	aPTT_s	$p < 0.001$
BE	$p < 0.05$	hemoglobin	$p < 0.05$
Lac	$p < 0.05$	ldh	$p = 0.127$
diuresis	$p < 0.001$	CEC duration	$p < 0.001$
aPPT_s	$p < 0.001$	adrenaline	$p < 0.001$

Table 4.7: Binary AKI case: Comparison between the first ten most important variables in RF, RF with RFE and corresponding p-value with GAMs

In the binary AKI case, we observe that only the variable ldh is in the list of the most important variables with a p-value greater than 0.05. We hypothesize that this non-significativity derives from the additivity constraints of the GAM model which, in this case, does not capture a more complex structure of interaction between the variables.

Similarly, in Table 4.8 we show the severe AKI case. In this case, the variable BE and clamping duration are in the list of the most important variables but they have a p-value greater than 0.05.

RF with all variables	GAM (p-value)	RF with subset via RFE	GAM (p-value)
creatinine	$p < 0.001$	creatinine	$p < 0.001$
CEC duration	$p < 0.001$	ldh	$p < 0.05$
basal creatinine	$p < 0.001$	platelets	$p < 0.05$
platelets	$p < 0.05$	basal creatinine	$p < 0.001$
lhd	$p < 0.05$	ethacrynic acid	$p < 0.05$
diuresis	$p < 0.001$	CEC duration	$p < 0.001$
ethacrynic acid	$p < 0.05$	aPTT_s	$p < 0.001$
aPPT_s	$p < 0.001$	hemoglobin	$p < 0.001$
BE	$p = 0.148$	PIM3	$p < 0.001$
Lac	$p < 0.001$	Clamping Duration	$p = 0.674$

Table 4.8: Severe AKI case: Comparison between the first ten most important variables in RF, RF with RFE and corresponding p-value with GAMs

Another meaningful mirror comparison consists to list the variables with a p-value <0.001 and see which of these fall within the top 10 most important variables of the classification with Random Forest.

In this, we compare the most important variables obtained with the Random Forests classification method and the variables with a p-value < 0.001 in the GAM model.

In Tables 4.9 and 4.10 we show the results of the binary case and the severe case respectively. In both cases, we notice that some variables like creatinine, basal creatinine, CEC duration, aPTT_s are inserted in each column.

GAM (p-value< 0.001)	RF	RF with RFE
weight		
age		
basal creatinine	basal creatinine	basal creatinine
PIM3		PIM3
CEC Duration	CEC Duration	CEC Duration
diuresis	diuresis	
fluid input		
fluid output		
creatinine	creatinine	creatinine
platelets	platelets	platelets
aPTT_s	aPTT_s	aPTT_s
heart rate		
adrenaline	adrenaline	adrenaline
ethacrynic acid		

Table 4.9: Binary AKI case: Comparison between most important variable in RF, RF with RFE and variables of GAM with $p - value < 0.001$.

It is important to notice that, if we consider the other variables inserted in the first ten most important variables, we observe that:

- in case of RF applied using all variables: all the other variables in the list have a p-value < 0.05 with the only exception of ldh and BE;
- in case of RF with RFE: all the other variables in the list have a p-value < 0.05 with the only exception of ldh and clamping duration.

GAM (p-value< 0.001**)**	**RF**	**RF with RFE**
basal creatinine	basal creatinine	basal creatinine
PIM3		PIM3
CEC duration	CEC duration	CEC duration
diuresis	diuresis	
fluid input		
fluid output		
creatinine	creatinine	creatinine
albumin		
hemoglobin		hemoglobin
aPPT_s	aPPT_s	aPPT_s
Lac	Lac	
adrenaline		
levosimendan		

Table 4.10: Severe AKI case: comparison between most important variable in RF, RF with RFE and variable of GAM with $p - value < 0.001$.

Chapter 5

Bayesian Network classification

5.1 Introduction

In addition to the previous approach with the GAMs, in this chapter, we explore the possibility of developing a model which offers the possibility of being more interpretable while keeping a reasonably high predictive performance using Bayesian Network (BN) classifiers. A BN classifier allows to bring out the role of the complex dependence structure of the variables that can influence the outcome more easily than the GAMs. BN classifiers are special types of Bayesian networks developed for classification problems. In particular, we focus on Bayesian network classifiers in the discrete domain. As described in Bielza and Larrañaga [2014], Bayesian network classifiers have several advantages. Among these, we point out the following:

- they provide an explicit, graphical, and interpretable representation of knowledge since they are an example of the probabilistic graphical model which can be associated with a graph;
- BN classifiers are computationally efficient algorithms, linear on the number of instances and linear, quadratic, or cubic (depending on model complexity) on the number of variables;
- both the binary and the multiclass cases can be treated with the same theoretical framework;
- many successful real-world applications have been reported in the literature with competitive performance against state-of-the-art classifiers.

Let $\mathbf{X}_s = (X_{s1}, ..., X_{sj}, ...X_{sD})$ be a vector of discrete random variables with $X_{sj} \in \Omega_{X_j} = \{1, 2, ...R_j\}$, and let C_s be a class (in our case the binary or multiclass AKI stage). Given a random sample $\mathcal{D} = (\mathbf{X}_1, C_1), ..., (\mathbf{X}_N, C_N)$ of size N, with a $(\mathbf{X}_s, C_s)$, drawn from the joint probability distribution $p(\mathbf{X}, C)$. The classification problem consists in finding a classification model from $\mathcal{D}$ able to assign a class to new data acquisitions given by the value of variables. A Bayes classifier assigns the most probable a posteriori (MAP) class to a given instance $\mathbf{x}$, that is:

$$\arg\max_c p(c|\mathbf{x}) = \arg\max_c p(\mathbf{x}, c)$$

. BN classifiers, as shown by Friedman et al. [1997], can approximate $p(\mathbf{x}, c)$ with a suitable factorization according to a Bayesian Network whose structure is a directed acyclic graph (DAG) with:

- vertices corresponding to random variables
- arcs encoding (in)dependences among variables.

In particular, each factor is a categorical distribution $p(x_i|\mathbf{pa}(x_i))$ or $p(c|\mathbf{pa}(c))$, where $\mathbf{pa}(\cdot)$ is a value of the set of variables which are parents of variable X_i or C according to the graphical structure.

Therefore we have that:

$$p(\mathbf{x}, c) = p(c|\mathbf{pa}(c)) \prod_{i=1}^{D} p(x_i|\mathbf{pa}(x_i))$$

The simplest and at the same time well-known Bayesian classifier is the naive Bayes one where all the predictor variables are conditionally independent. In particular, all the X_i are conditionally independent given C. In figure 5.1 we show an example of a naive Bayes structure described by the following relation:

$$p(c, \mathbf{x}) = p(c)p(x_1|c)p(x_2|c)p(x_3|c)p(x_4|c)p(x_5|c)p(x_6|c)$$

.

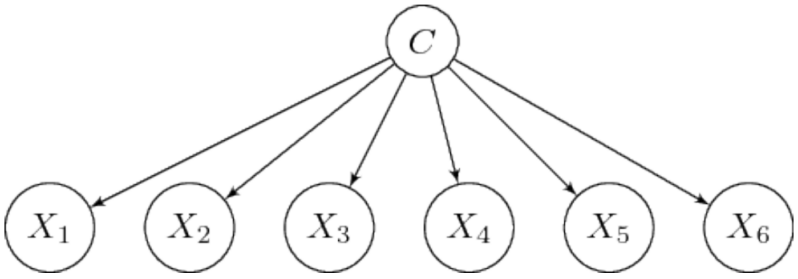

Figure 5.1: Example of naive Bayes structure schema.

In most cases, the naive Bayes approach is a fairly crude approximation. More flexible and hence more realistic models are One-Dependence Estimators (ODEs). ODEs are similar to naive Bayes except for the fact that, by relaxing the basic global conditional independence constraint among all the X's, each predictor variable is allowed to depend on at most another predictor in addition to the class c.

Among several possible implementations of ODEs, the tree-augmented network (TAN) (see Friedman et al. [1997]) maintains the original predictor variables and models relationships of at most order 1 among the variables. An example of TAN structure is depicted in Figure 5.2 and described by the following equation:

$$p(c, x) = p(c)p(x_1|c, x_2)p(x2|c, x_3)p(x_3|c, x_4)p(x_4|c)p(x_5|c, x_4)p(x_6|c, x_5)$$

.

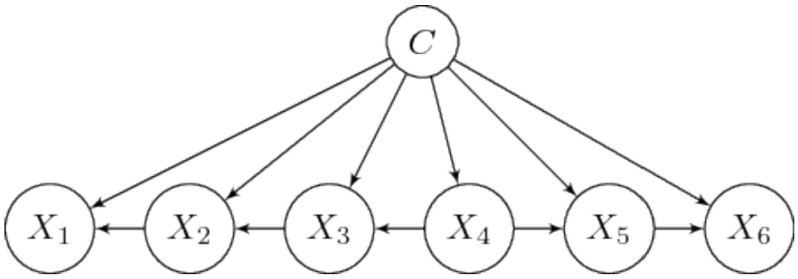

Figure 5.2: Example of TAN structure schema.

Friedman et al. [1997] developed TAN on the basis of a procedure for learning edges adapting from a method reported by Chow and Liu (see Chow and Liu [1968]).

Apart from learning the network structure, the probabilities $p(x_i|\mathbf{pa}(x_i))$ are estimated from $\mathcal{D}$ by standard methods like maximum likelihood or Bayesian estimation. For discrete variables X_i and $\mathbf{pa}(X_i)$, Bayesian estimation can be obtained in closed form by assuming symmetric Dirichlet priors over the underlying vector of probability masses denoted with θ_{ij}. With all Dirichlet hyperparameters equal to α,

$$\hat{\theta}_{ijk} = \frac{N_{ijk} + \alpha}{N_{\cdot j \cdot} + r_i \alpha},$$

where N_{ijk} is the number of instances in $\mathcal{D}$ such that $X_i = k$ and, with a slight abuse of notation, $\mathbf{pa}(x_i) = j$, corresponding to the j-th possible instantiation of $\mathbf{pa}(x_i)$, $N_{\cdot j \cdot}$ is the number of instances in which $\mathbf{pa}(x_i) = j$, while r_i is the cardinality of X_i. If in the previous equation, $\alpha = 0$, which corresponds to the limit of Bayesian estimation with α tends to zero, we have the case of maximum likelihood estimation of θ_{ijk}.

To apply the BN classifier to our dataset we use the *bnclassify* R package developed by Mihaljevi'c et al. [2018]. In order to use this, we have to make some preliminary preprocessing of the available data in order to fit within the framework of a discrete Bayesian network.

In particular we:

- discretize all the continuous predictor variables using a "quantile-based" approach deciding a number of intervals n_{int} able to maximize the classification performance;
- Use for each continuous variable a number of intervals equal to n_{int} or the maximum number of possible intervals until to n_{int}
- Split our data in learning and test (70% for learning and 30% for test) preserving the percentages of each class in train and test sets.
- Involve all the variables of the model.
- Use the TAN Bayes structure with $\alpha = 0.1$; this approach showed the best results in terms of classification performance.

The discretization was made using the *RoughSet R package* implemented by Riza et al. [2014]. We have also implemented alternative discretization approaches like intervals with equal width or using k-means clustering obtaining similar or inferior results.

Instead of defining a priori a constrained structure of the network topology we have considered leaving the structure completely unspecified and attempted to learn the structure (see Nagarajan et al. [2013]) and making classification from this DAG. Although potentially less affected by constraints such as TANs, this approach gave a much poorer predictive performance and, for this reason, we do not report them here.

5.2 Binary classification case with BN

In this section we show the results obtained by applying BN classifiers to binary classes in the following cases:

- binary AKI;
- severe AKI.

For both cases, we obtain the best results with a maximum number of interval discretizations $n_{int} = 17$ as described in Figure 5.3.

The results obtained in the binary AKI case are shown in Table 5.1. In particular, considering the reference class AKI 1, we obtain:

- $AUC - ROC = 0.90$
- $sensitivity = 0.72$
- $specificity = 0.90$

Instead, the results obtained in severe AKI case are shown in Table 5.2. In particular, considering the reference class AKI 1, we obtain:

- $AUC - ROC = 0.97$
- $sensitivity = 0.82$
- $specificity = 0.94$

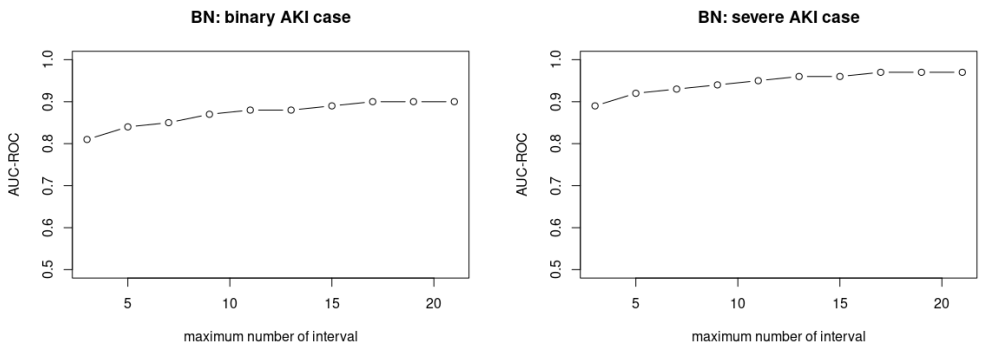

Figure 5.3: Plot of AUC-ROC vs max number of intervals: binary and severe AKI case

	48h binary AKI score using BNs	CI
SENS	0.72	0.69-0.75
SPEC	0.90	0.89-0.91
TP	530	-
FP	323	-
TN	2797	-
FN	205	-
AUC-ROC	0.90	0.88-0.92

Table 5.1: AUC-ROC of BN model (binary AKI case with all variables)

In Figures 5.4 and 5.5 we show the corresponding DAG respectively for binary and severe AKI cases. Considering the TAN constraints the connection between the variables seems reasonable (many variables connected with weight, PIM3, pressure, etc...) and similar for both binary and severe AKI cases. It is important to note that these DAGs have the limitations of deriving from an ODEs Bayesian structure and only part of them can comply with the dynamics expected by biological functioning.

	48h severe AKI score using BNs	CI
SENS	0.82	0.79-0.85
SPEC	0.94	0.93-0.95
TP	389	-
FP	207	-
TN	3175	-
FN	84	-
AUC-ROC	0.97	0.96-0.98

Table 5.2: AUC-ROC of BN model (severe AKI case with all variables)

5.3 Multiclass case

In this section, we show the results obtained in the multiclass case. For both mode and max cases, we obtain the best results with a maximum number of interval discretizations $n_{int} = 27$ as described in Figures 5.6 and 5.7.

In Table 5.3 we show the confusion matrix in the mode AKI case. As a summary score, we highlight:

- Accuracy= 0.92
- kappa= 0.73

Instead, in Table 5.4 we depict the mode AKI case. In this case, we have:

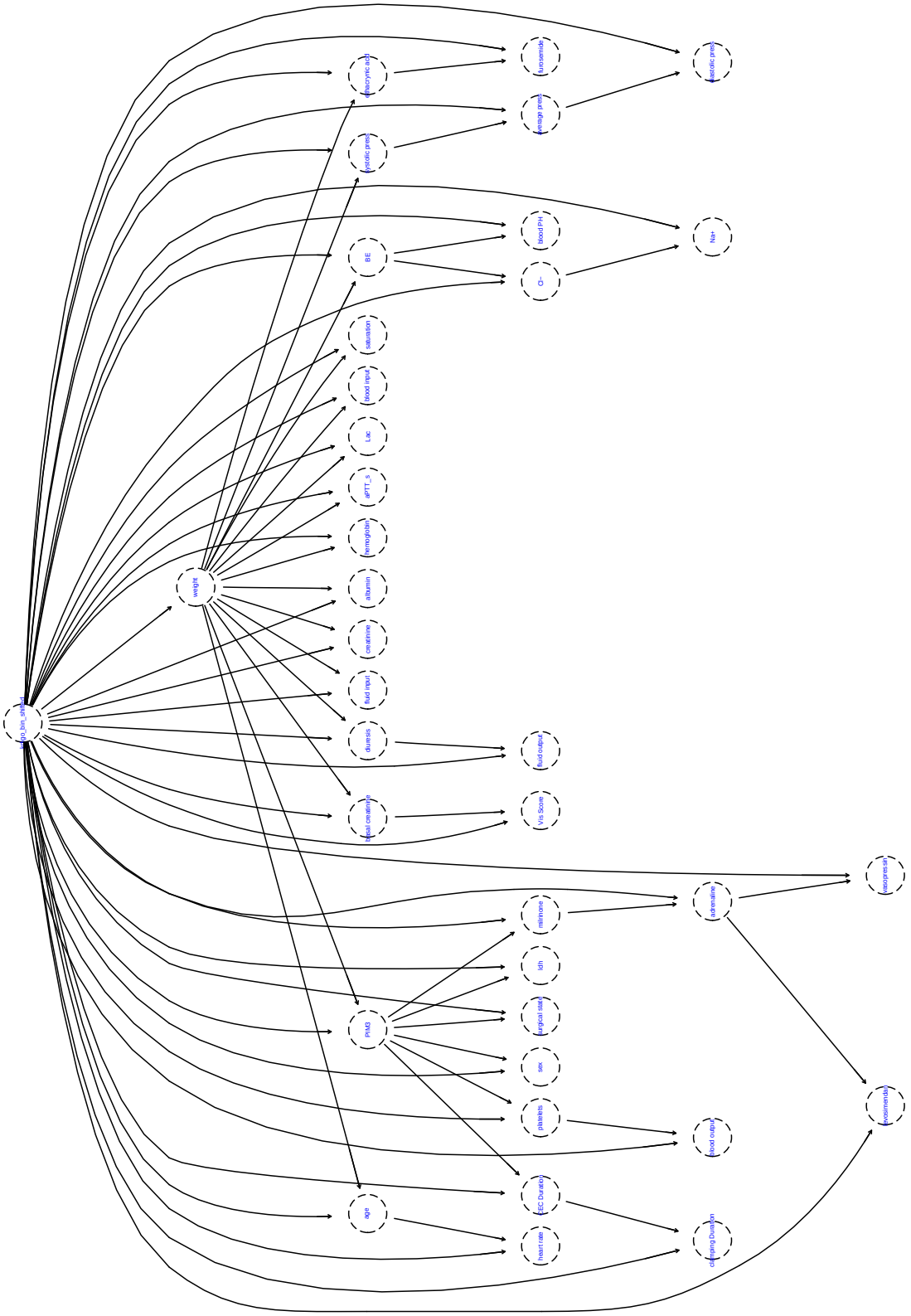

Figure 5.4: DAG binary AKI case

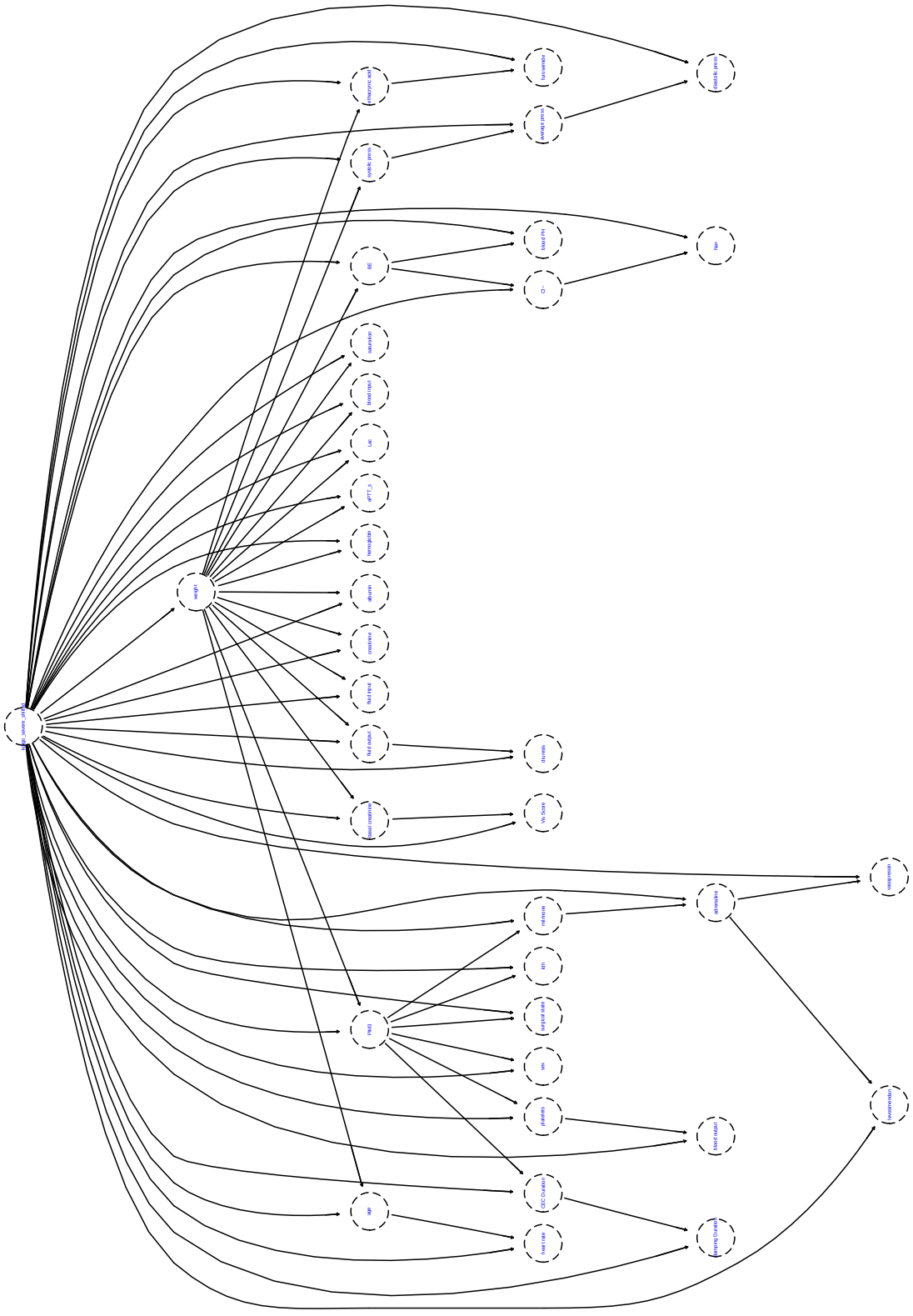

Figure 5.5: DAG severe AKI case

- Accuracy= 0.88
- kappa= 0.64

If we look at the confusion matrix reported in Tables 5.3 and 5.4, we observe the difficulty in correctly classifying the class of AKI 1. This problem is related to the fact that AKI 1 is less severe and therefore more easily confused with the class of AKI 0 (that is, with the status of a healthy patient).

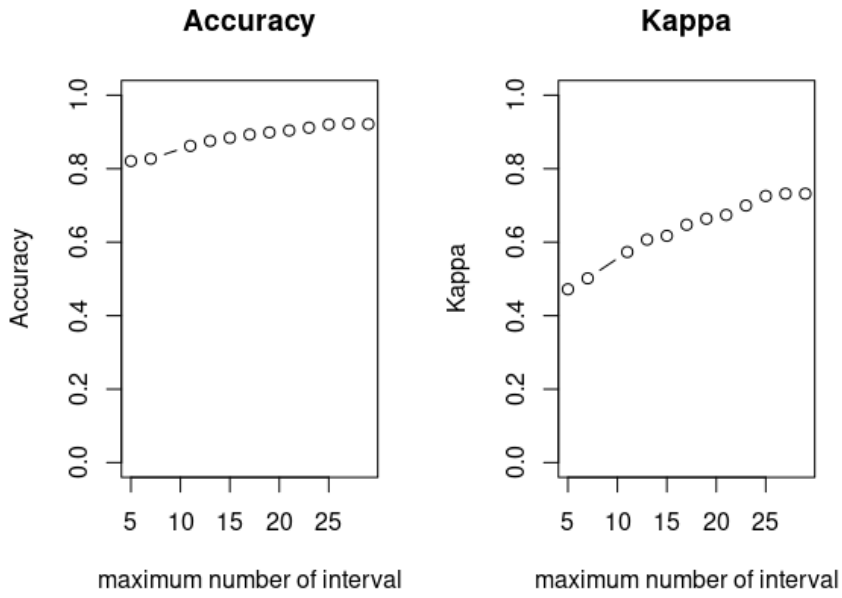

Figure 5.6: Plot of accuracy and kappa vs max number of intervals: mode AKI case

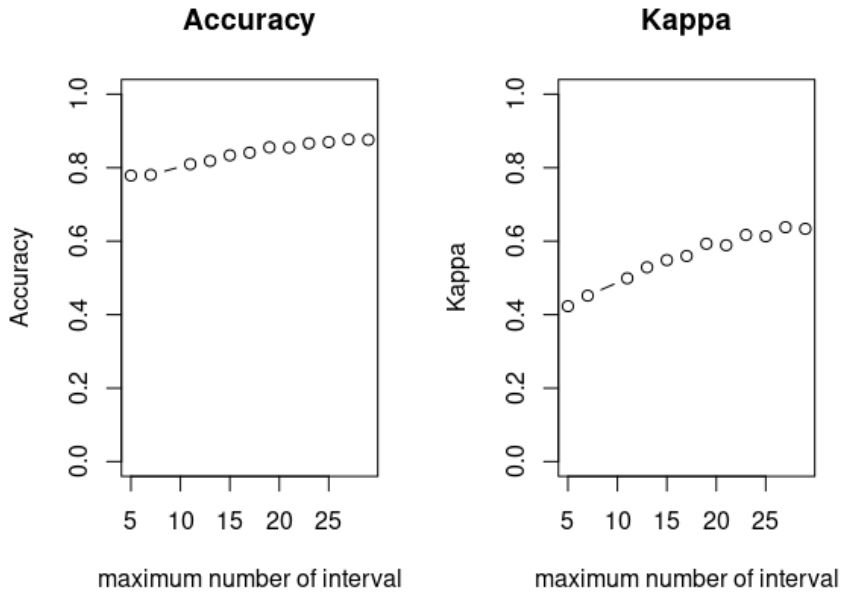

Figure 5.7: Plot of accuracy and kappa vs max number of intervals: max AKI case

Mode AKI confusion matrix				
	Observed			
Prediction	0	1	2	3
0	3097	62	26	20
1	50	98	7	1
2	54	2	132	10
3	54	5	7	229

Table 5.3: Confusion matrix: mode AKI case using BN approach

Max AKI confusion matrix				
	Observed			
Prediction	0	1	2	3
0	2891	123	41	25
1	105	118	10	5
2	74	13	133	19
3	50	7	2	237

Table 5.4: Confusion matrix: max AKI case using BN approach

In Figures 5.8 and 5.9 we show the corresponding DAG respectively for cases of AKI mode and AKI max.

Considering the TAN constraints, the connection between the variables seems reasonable (many variables connected with weight, PIM3, pressure, etc...) and similar to each other and to the previous binary cases.

5.4 Variable selection in the BN case using MXM package

Also in the case of BN, as in the case of classification with the RF method, we try to identify a subset of variables that allow fitting a less complex DAG leading to a BN with similar predictive performance.

The basic idea is that a graph with fewer variables can be interpreted more simply from a medical point of view.

In order to do this we use the *MXM* R Package.

As explained by Lagani et al. [2017], MXM, short for the latin 'Mens ex Machina' [(Human) Mind from the Machine], is a collection of utility functions for variable selection especially developed in the case of BNs.

In particular, we use the Max-Min Parents and Children algorithm (MMPC) algorithm. It is a constraint-based feature selection algorithm that assumes a Bayesian Network for all observed variables. Parents and Children refer to the fact that the algorithm identifies the parents and children of the variable of interest.

MMPC algorithm performs multiple conditional independence tests and progressively excludes irrelevant and redundant variables.

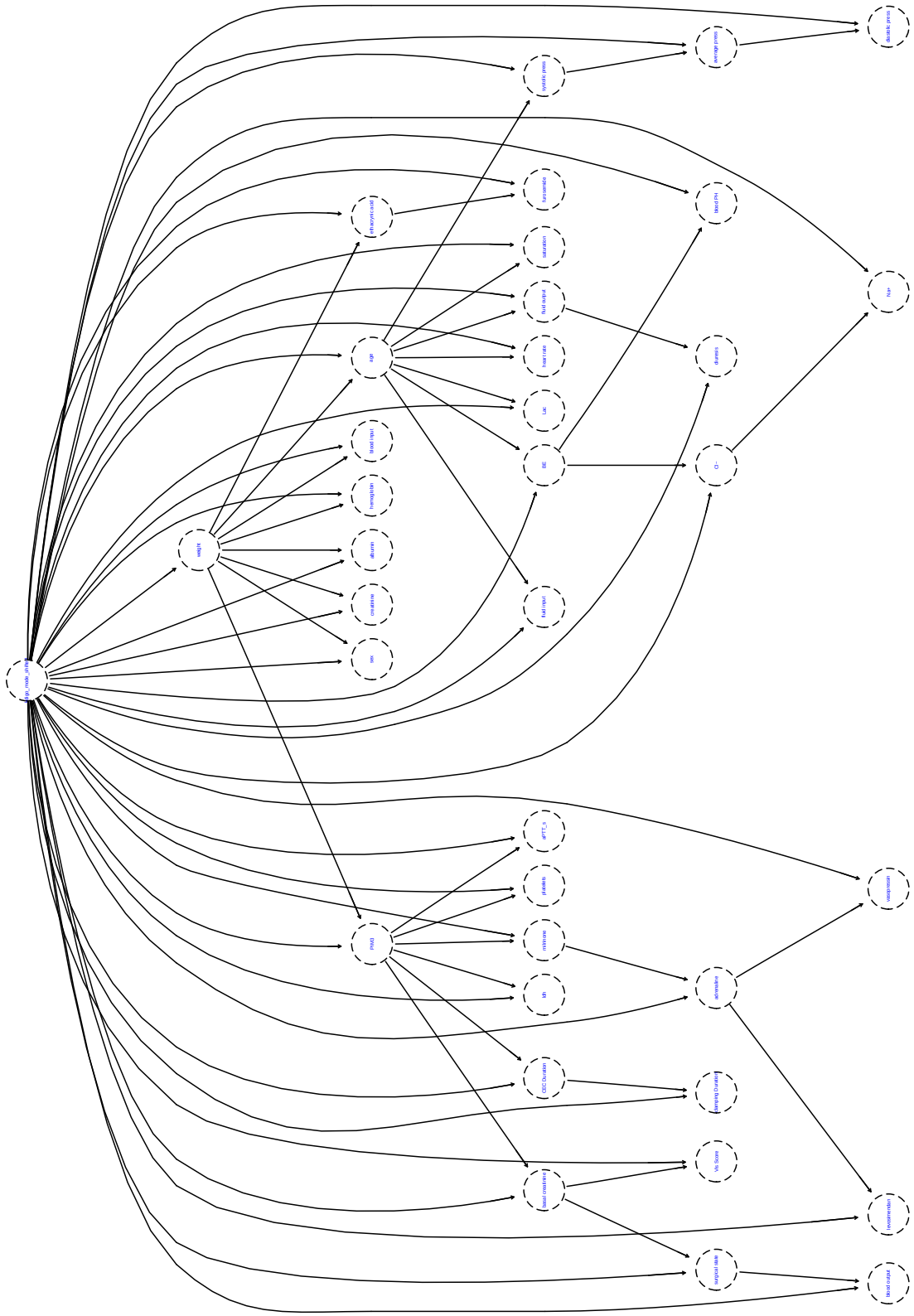

Figure 5.8: DAG mode AKI case

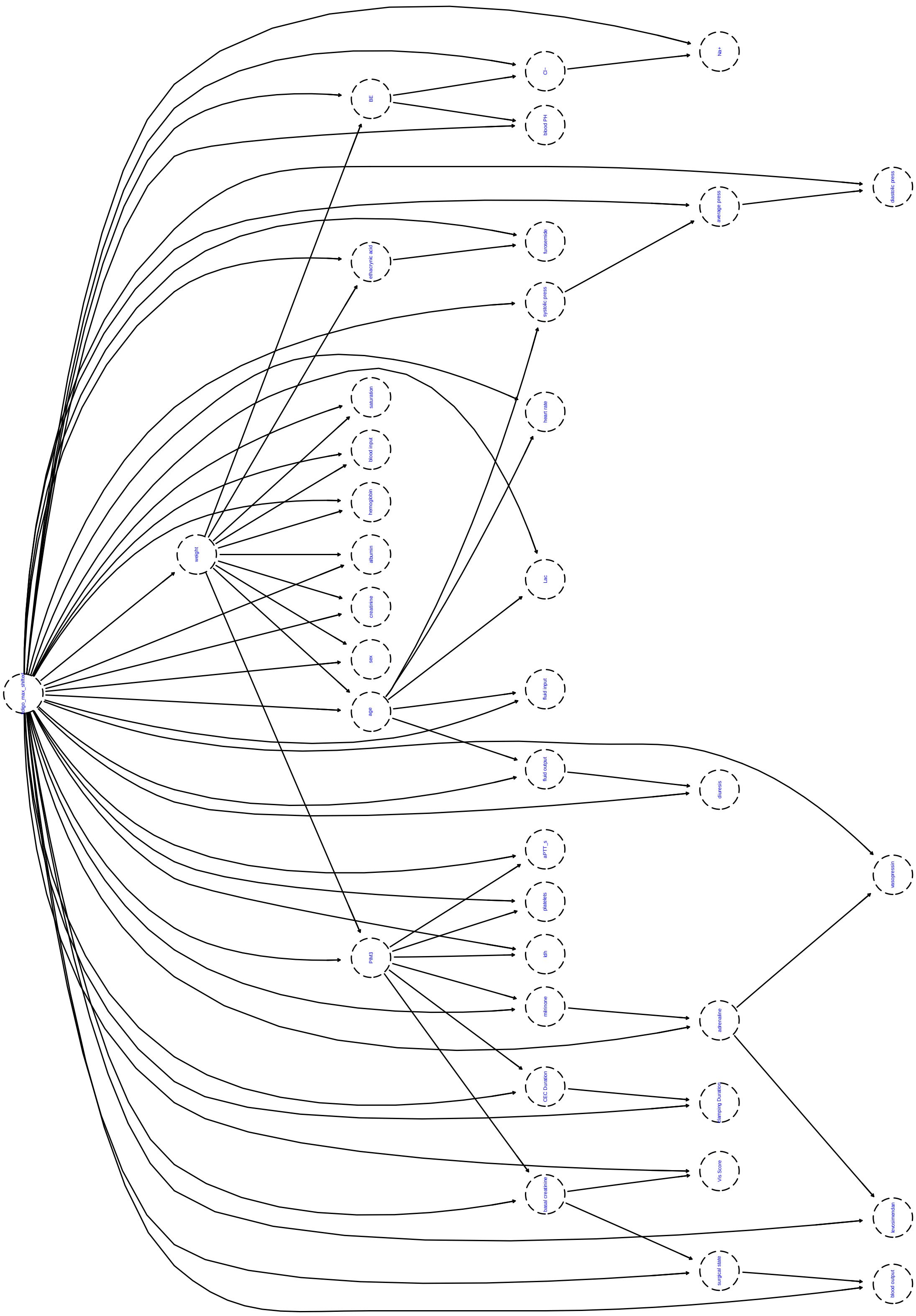

Figure 5.9: DAG max AKI case

The final variables that have "survived" through all those elimination stages are the MMPC output signature.

The MMPC algorithm discovers the subset of variables using a two-phase procedure:

- in phase 1, the forward phase, MMPC includes in a set of candidates, denoted as *CPC* (candidate parents and children) the variable with the highest univariate association with the target variable T. MMPC chooses to include next into the *CPC* the variable that exhibits the maximum association with T conditioned on the subset of *CPC*;
- in phase 2, the backward phase, MMPC examines whether each variable of *CPC* can be d-separated from T by conditioning on all possible subsets of *CPC* and, in case, removes it.

For details of the algorithm, we refer to Brown et al. [2004] and Tsamardinos et al. [2003].
We proceed with the following steps starting from the discretized dataset:

- identify the subset of the variables using the MMPC algorithm applied to the training dataset;
- apply the BN classifier using the selected subset of variables in the same way as described in paragraph 5.1.

.

For completeness, we report that we also tested the use of another variables selection algorithm available in MXM package with the acronyms of IAMB (Incremental Association Markov Blanket) but, despite the promising performances alleged by the authors, it proved itself less performing than MMCP.

In the following sections, we show the results obtained in the binary and multiclass cases.

5.4.1 BN Binary case using a subset of variables identified with MMPC

In this paragraph, we show the results obtained by applying BN classifiers to binary class and the selected subset of variables using MXM package in the following case:

- binary AKI;
- severe AKI.

. For both cases, the number of selected variables is 25.

The list of these variables is reported in Table 5.5. We observe that a large number of variables (specifically 19) are selected in both cases.

binary AKI	severe AKI
creatinine	creatinine
basal creatinine	basal creatinine
diuresis	weight
adrenaline	PIM3
weight	diuresis
BE	ethacrynic acid
PIM3	ldh
furosemide	Lac
hemoglobin	hemoglobin
platelets	blood PH
Cl-	systolic press
age	platelets
Lac	age
blood input	adrenaline
fluid output	BE
ethacrynic acid	CEC Duration
blood output	aPTT_s
albumin	furosemide
levosimendan	blood input
systolic press	Vis Score
vasopressin	fluid output
clamping Duration	albumin
Na+	Cl-
blood PH	fluid input
Vis Score	levosimendan

Table 5.5: List of variables selected using MXM sorted by highest statistical significance (binary and severe AKI cases)

In Tables 5.6 and 5.7, we report the results of performance for binary and severe AKI case evaluated using the test dataset.

From our point of view it is important to notice that we obtain the same AUC-ROC obtained using all the model variables despite using fewer variables.

In fact, the results are the following:

- binary case: AUC-ROC= 0.90 (0.89-0.91);
- severe case: AUC-ROC= 0.97 (0.96-0.98).

In Figures 5.10 and 5.11 we show the corresponding DAGs. With regard to DAGs obtained in this case, we observe that in both cases there is a direct link between KDIGO and creatinine. Previously this connection was made by first passing through the weight variable. This direct connection has an interpretation closer to what is expected from the point of view of the biological phenomenon.

	48h binary AKI score using BNs with MXM subset	**CI**
SENS	0.73	0.70-0.76
SPEC	0.89	0.88-0.90
TP	536	-
FP	342	-
TN	2778	-
FN	199	-
AUC-ROC	0.90	0.88-0.92

Table 5.6: AUC-ROC of BN model (binary AKI case with MXM subset of variables)

	48h severe AKI score using BNs with MXM subset	**CI**
SENS	0.80	0.76-0.83
SPEC	0.95	0.94-0.96
TP	378	-
FP	175	-
TN	3207	-
FN	95	-
AUC-ROC	0.97	0.96-0.98

Table 5.7: AUC-ROC of BN model (severe AKI case with MXM subset of variables)

5.4.2 Multiclass case

Here, we show the results obtained by applying BN classifiers to the multiclass case with a subset of variables selected using the MMPC algorithm implemented in MXM package.

As before, we have the following case:

- mode AKI;
- max AKI.

. The number of variables selected by MXM was 22 for the mode AKI and 26 for the max AKI.

The list of these variables is reported in Table 5.8. Also, in this case, we observe that a large number of variables (specifically 21) are selected in both cases. Moreover, we notice that the variables involved are substantially the same as in the binary case.

In Table 5.9 we show the confusion matrix in the mode AKI case. As a summary score, we highlight:

- Accuracy= 0.92
- kappa= 0.72

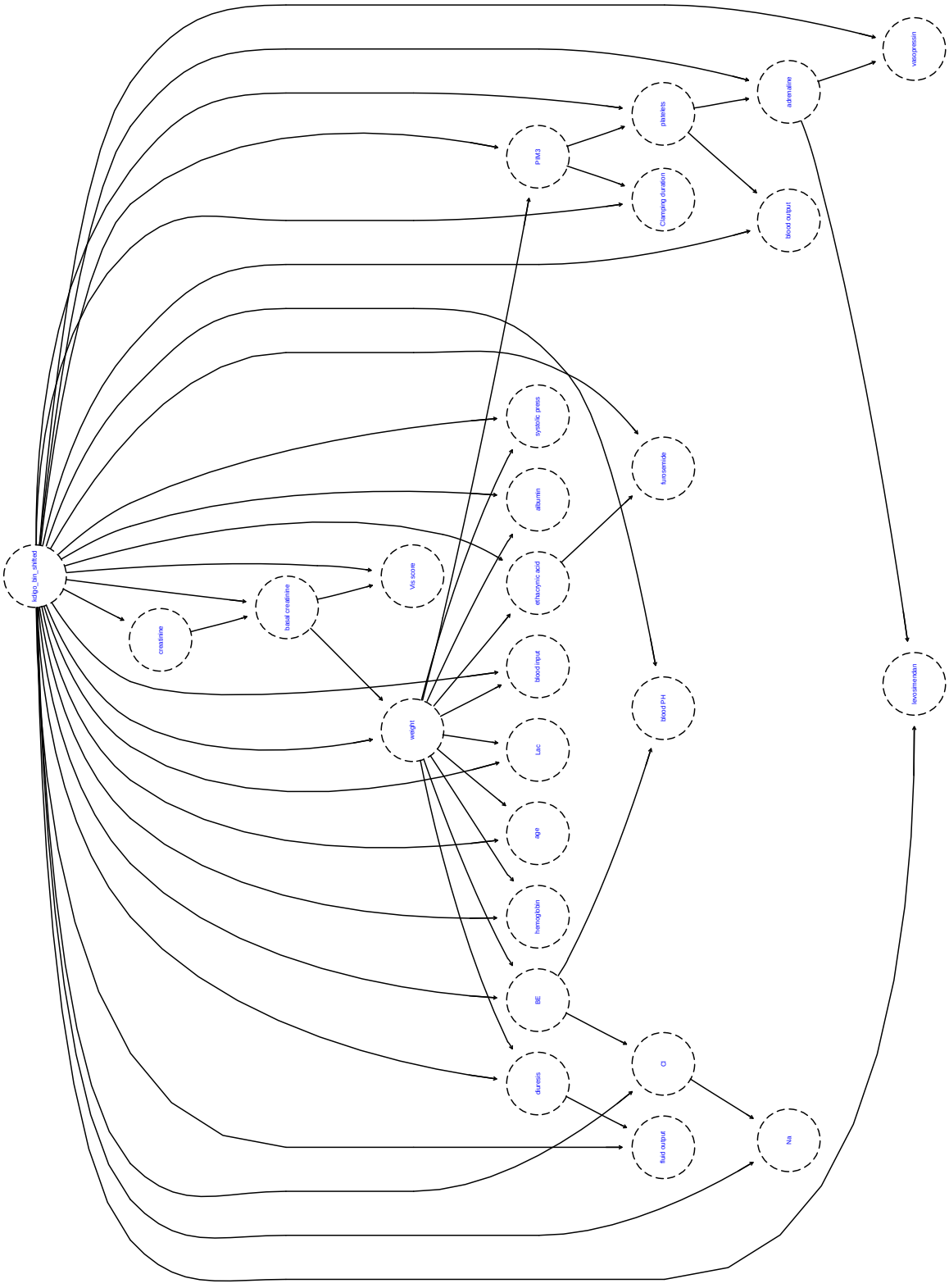

Figure 5.10: DAG binary AKI case (MXM subset of variables

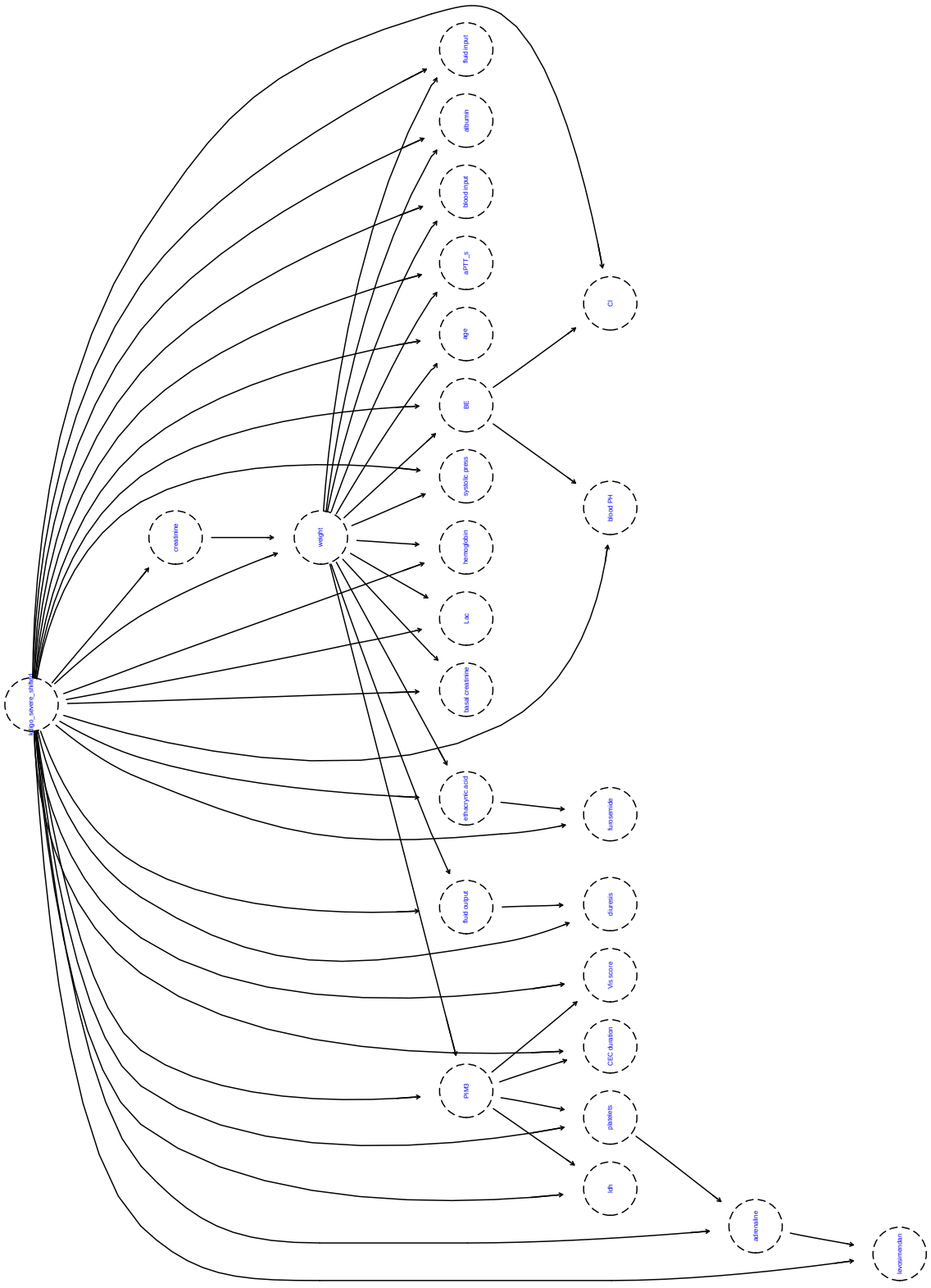

Figure 5.11: DAG severe AKI case (MXM subset of variables

mode AKI	max AKI
creatinine	creatinine
basal creatinine	basal creatinine
PIM3	adrenaline
adrenaline	diuresis
diuresis	PIM3
weight	weight
platelets	BE
BE	ethacrynic acid
ldh	platelets
ethacrynic acid	hemoglobin
albumin	systolic press
hemoglobin	ldh
aPTT_s	age
systolic press	aPTT_s
age	Vis Score
Vis Score	blood input
CEC Duration	albumin
levosimendan	levosimendan
blood input	Cl-
Cl-	fluid input
blood output	furosemide
fluid output	CEC Duration
	vasopressin
	Lac
	Na+
	blood output

Table 5.8: List of variables selected using MXM sorted by highest statistical significance (mode and max AKI cases)

In Table 5.10 we depict the max AKI case. In this case, we have:

- Accuracy= 0.87
- kappa= 0.63

Also in this case we observe that the final results are substantially the same as those obtained using all the variables.

According to "Occam's razor" principle, the advantage of a simpler model compared to a more complex one is clear, especially as already mentioned, in interacting with doctors.

Mode AKI confusion matrix				
	Observed			
Prediction	0	1	2	3
0	3050	58	20	18
1	84	101	5	1
2	65	5	142	9
3	56	2	5	232

Table 5.9: Confusion matrix: mode AKI case (BN approach with MXM subset of variables)

In Figures 5.12 and 5.13 we show the corresponding DAG respectively for cases of AKI mode and AKI max.

As for the binary, we observe a direct link between AKI KDIGO and creatinine.

It is confirmed that the connection between the variables seems reasonable (e.g. many variables are connected with weight or PIM3) and similar to each other and to the previous binary cases.

Max AKI confusion matrix				
	Observed			
Prediction	0	1	2	3
0	2861	121	32	35
1	148	124	10	3
2	62	11	139	11
3	49	5	5	237

Table 5.10: Confusion matrix: max AKI case (BN approach with MXM subset of variables)

5.5 Overview of comparative performance of alternative classifiers

In this section, we insert a comparison of the results of all the previous methods.

In order to do this, we insert in Table 5.11 a summary of the results. In particular, we insert:

- binary case: AUC-ROC value;
- multiclass case: accuracy and kappa.

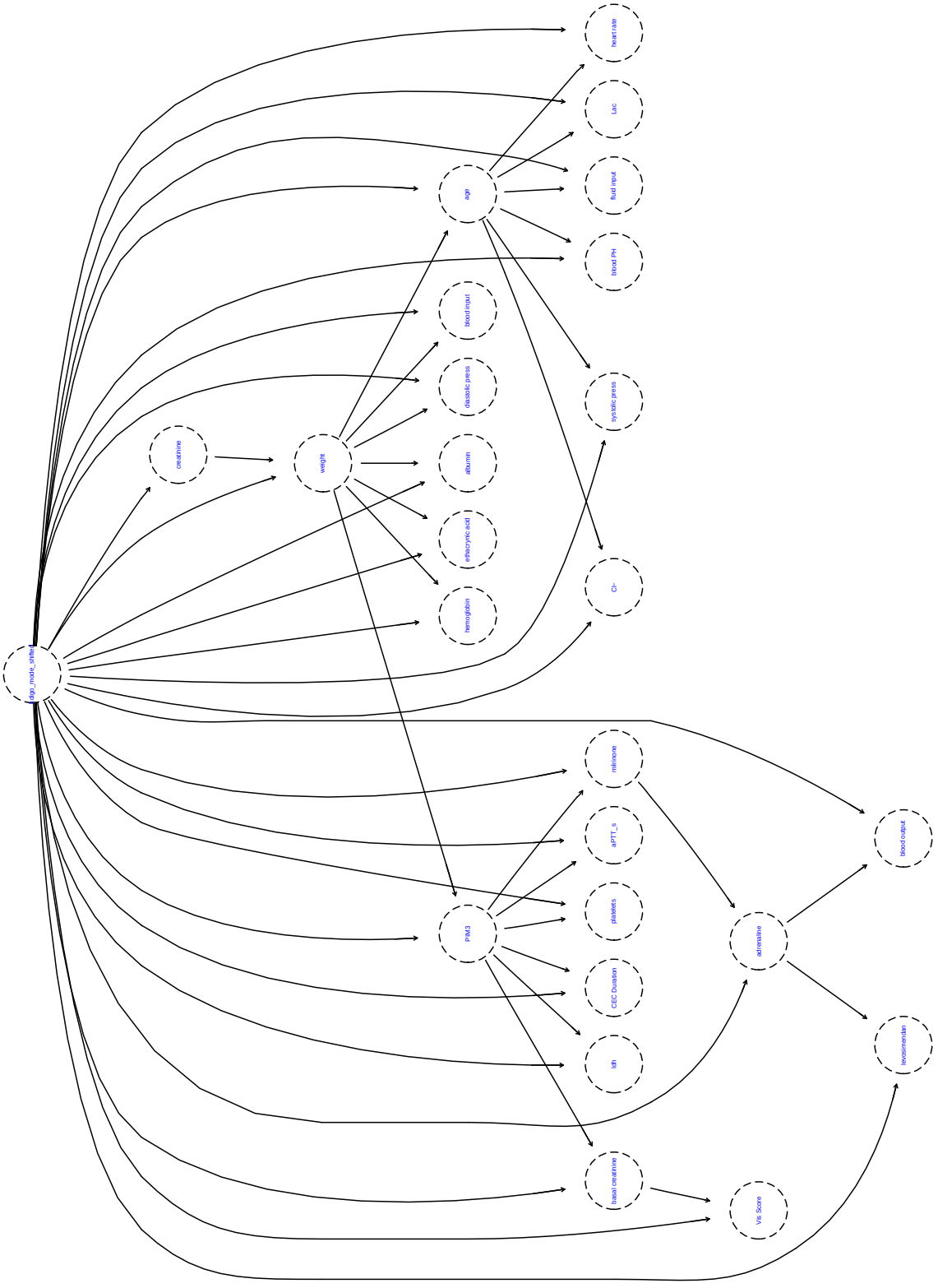

Figure 5.12: DAG mode AKI case (MXM subset of variables)

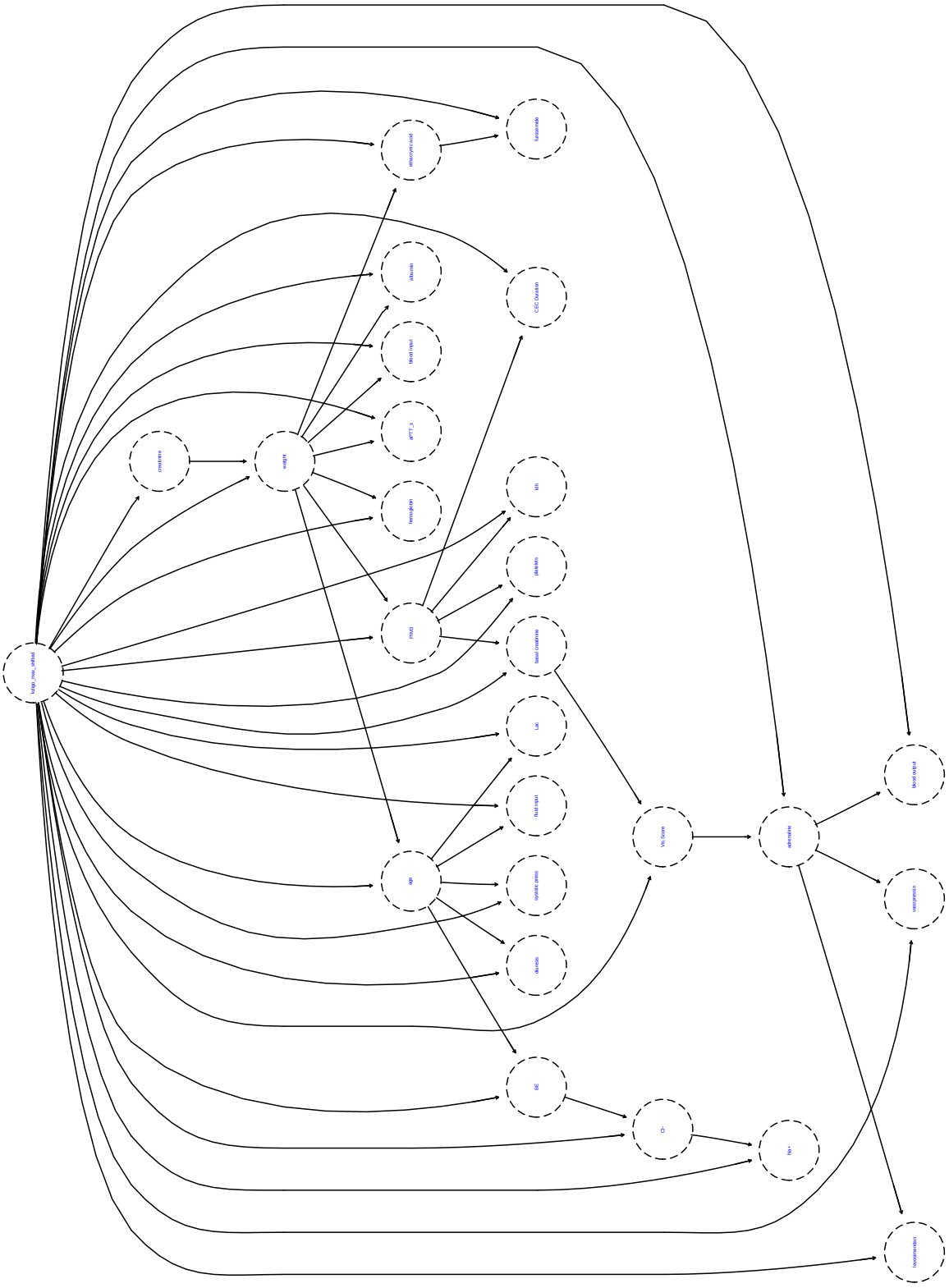

Figure 5.13: DAG max AKI case (MXM subset of variables)

We observe that the best results are obtained using the RF classification method.

In particular, the RF using a subset of variables obtained using RFE is preferable because it uses fewer variables.

Despite this, we think that the BN approach is able to provide a model that is interpretable and provides, at the same time, good results in terms of performance.

In the binary class case, all the results with an AUC-ROC greater than 0.9 are considered very good from the point of view of doctors.

These results obtained with different methods suggest continuing this analysis with a larger dataset and, subsequently, evaluating the use of these results in a prospective way.

Multiclass results seem to suggest evaluating a system that makes a classification using only three classes merging AKI 0 and AKI 1 classes in the "not severe AKI" class.

Moreover, we chose the use of multiple classification techniques in an attempt of searching for a better understanding of the phenomenon beyond the achievement of the best result in terms of performance.

Summary of results obtained with different methods					
	RF (all var)	**RF using RFE**	**GAM (all var)**	**BN (all var)**	**BN (subset MXM)**
AUC bin AKI	0.93 (0.92-0.94)	0.95 (0.94-0.96)	0.87 (0.85-0.89)	0.90 (0.88-0.92)	0.90 (0.88-0.92)
AUC severe AKI	0.99 (0.98-1)	0.98 (0.97-0.99)	0.94 (0.92-0.96)	0.97 (0.96-0.98)	0.97 (0.96-0.98)
Accuracy and Kappa Max AKI	acc= 0.92 (0.91-0.93) k= 0.71	acc= 0.93 (0.92-0.93) k= 0.76	acc= 0.87 (0.86, 0.89) k= 0.54	acc= 0.88 (0.87-0.89) k= 0.64	acc= 0.87 (0.86-0.88) k= 0.63
Accuracy and Kappa Mode AKI	acc= 0.95 (0.94-0.96) k= 0.80	acc= 0.96 (0.95-0.97) k= 0.85	acc= 0.90 (0.89, 0.91) k= 0.59	acc= 0.92 (0.91-0.93) k= 0.73	acc= 0.92 (0.91-0.92) k= 0.72

Table 5.11: Summary of results obtained using respectively: RF, RF with RFE, GAM, BN, BN with MXM.

Chapter 6

Some insights on the interpretation of RF for AKI prediction

6.1 Interpretable Machine Learning approach

Recent developments in Machine Learning algorithms have shown great potential for improving the accuracy of predictions. However, this predictive ability is often not supported by the possibility of interpreting the developed model.

Hence, there has been a growing interest in developing methods capable of interpreting black box models (see Molnar [2022]). In fact, the lack of interpretability limits the application of the model in academia, industry, and, particularly, the medical field.

Methods capable of making a black box more explainable overcome this limitation and, for this reason, they are of great interest. Among the various possible developments in the literature we can distinguish between the following approach:

- model-agnostic vs model-specific interpretation;
- global vs local model.

The model-agostic approach differs from the model-specific by the fact that it is model-independent: it separates the explanations from the machine learning model. As described by Ribeiro et al. [2016b] the great advantage of model-agnostic interpretation methods over model-specific ones is their flexibility. Machine learning developers are free to use any machine learning model they like when the interpretation methods can be applied to any model.

Local interpretation methods explain individual predictions for example by replacing the complex model with a locally interpretable surrogate model. In Figure 6.1 we insert an image that visually describes this method.

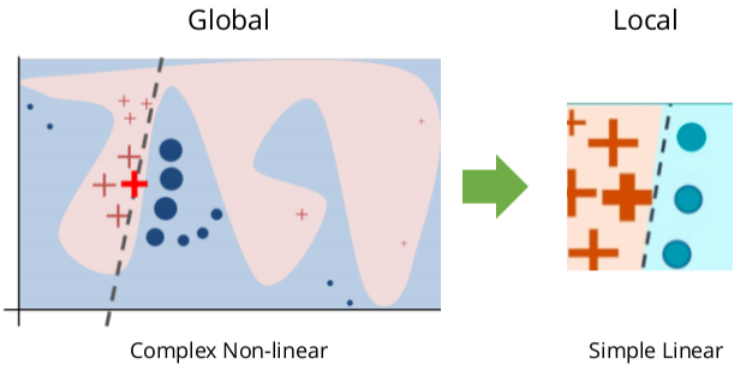

Figure 6.1: Local interpretation methods explained visually

Instead, global interpretation methods replace the complex model with a simpler and globally interpretable model.

In this chapter, we insert a further analysis of the results obtained with RFs in order to increase the explainability of these results on behalf of medical experts.

In the next paragraphs we will focus on:

- Partial Dependence Plots (PDPs) of the most important variables;
- Decision Tree Surrogate (DTS) approach.

PDPs are a useful tool for black-box model visualization. DTS is a global surrogate model developed using decision trees trained to approximate the predictions of a black box model. Both PDPs and DTS are model-agnostic interpretation methods since they separate the "explanations" from the machine learning method.

6.1.1 Partial Dependence Plot

The Partial Dependence plot shows the change in the average predicted value as the specified variable, or set of variables, varies over their marginal distribution.

This idea was originated from Friedman [2001] and was extended by Goldstein et al. [2015].

Let $L \subset \{1, ..., D\}$ and let M be the complement set of L where L and M index subsets of predictors. Then the partial dependence function of f on $\mathbf{x}_L$ is given by

$$f_L(\mathbf{x}_L) = \mathbb{E}_{\mathbf{x_M}}[f(\mathbf{x}_L, \mathbf{x}_M)] = \int f(\mathbf{x}_L, \mathbf{x}_M) dP(\mathbf{x}_M)$$

Each subset of predictors L has its own partial dependence function f_L, which gives the average value of f when $\mathbf{x}_L$ is fixed and $\mathbf{x}_M$ varies over its marginal distribution $dP(\mathbf{x}_M)$. Since neither the true f nor $dP(\mathbf{x}_M)$ are known, we estimate previous equation by computing:

$$\hat{f}_L = \frac{1}{N} \sum_{i=1}^{N} \hat{f}(\mathbf{x}_L, \mathbf{x}_{Mi})$$

where $\mathbf{x}_{M1}, ..., \mathbf{x}_{MN}$ are the values of $\mathbf{x}_M$ observed in the training data.

Partial dependence works by marginalizing the machine learning model output over the distribution of the variables (also denoted in the literature as features) in set M, so that the function shows the relationship between the variables in set L we are interested in, and the predicted outcome. By marginalizing over the other variables, we get a function that depends only on variables in L, interactions with other variables included.

The partial function tells us for a given value(s) of variables x_L what is the average marginal effect on the prediction. This is a visualization tool because a set of N pairs $(\mathbf{x}_{Ll}, \hat{\mathbf{f}}_{Ll})_{l=1}^{N}$ is obtained when $\hat{f}_L$ is evaluated at the $\mathbf{x}_L$ observed. The resulting graphic is called a Partial dependence plot and obviously, it has the limitation of being displayed only for 1 or 2 dimensions. As reported in Goldstein et al. [2015] several supervised learning models have been better-understood thanks to PDPs. Another limitation is related to the fact that it ignores possible feature interactions and captures only the marginal effect of the variable: a variable could be very important but the partial dependence plot could be flat since the variable affects the prediction mainly through interactions with other variables.

The partial dependence plot is an agnostic and global method; in fact:

- it is applicable regardless of the classification model used;
- it considers all instances and gives a statement about the global relationship of a feature with the predicted outcome.

6.1.2 Decision Tree surrogate approach

As described in Molnar [2022] and Craven and Shavlik [1995], a way to make models more interpretable is to replace the "black box" model with a simpler (also known as a surrogate) model such as a decision tree. In figure 6.2 we depict visually this idea. It can be described by the following steps:

- apply the original model and get predictions;
- choose an interpretable "white box" model (in our case, a decision tree);
- train the interpretable model on the original dataset and its predictions;
- measure how well the surrogate model agrees with the prediction of the black box model.

In *iml* R package Molnar et al. [2018], R^2 is used in order to measure how well the decision tree approximates the "black box" model. It is calculated in the following way:

$$R^2 = 1 - \frac{\sum_{s=1}^{n}(\hat{y}_*^{(s)} - \hat{y}^{(s)})^2}{\sum_{s=1}^{n}(\hat{y}^{(s)} - \bar{\hat{y}})^2}$$

where $\hat{y}_*^{(s)}$ is the prediction for the s-th instance of the surrogate model. For the multi-class case, R^2 contains one measure per class.

It is essential to notice that the surrogate model makes statements about the model and not about the "real world". The interpretation of the surrogate model is not good if the black box model is bad, because then the black box model itself is not useful.

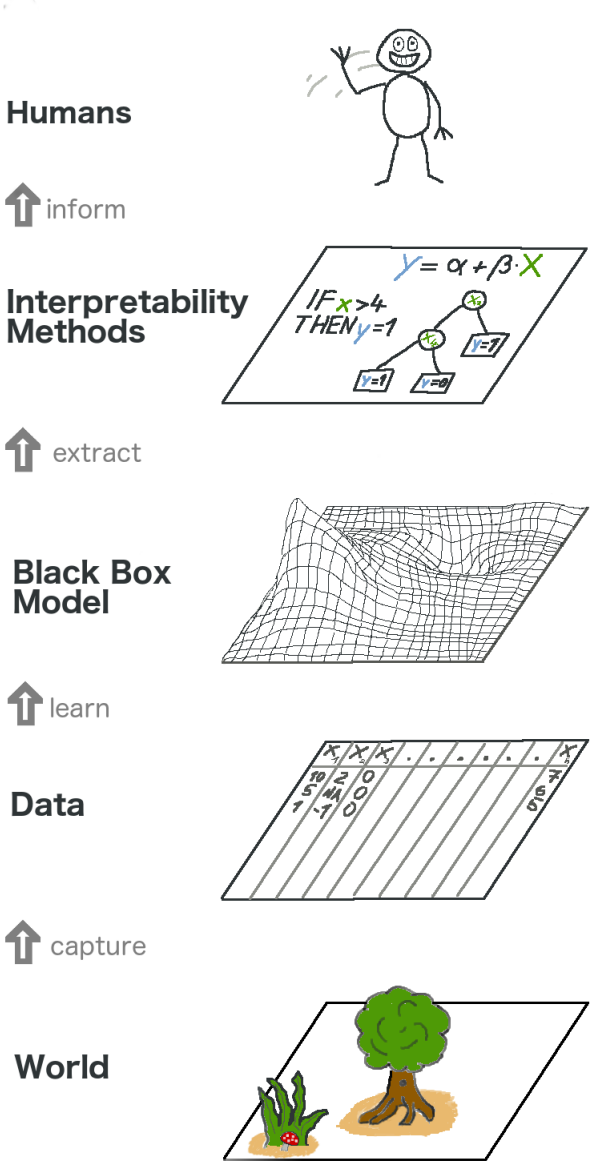

Figure 6.2: Surrogate model schema (image from Molnar [2022]).

As already said in the introduction of this chapter, it is important to notice that, parallel to the global approaches, there are local approaches. Local agnostic surrogate models are interpretable models that are used to explain individual predictions of black box machine learning models. A paper of Ribeiro et al. [2016a] describes the idea of a local interpretable model-agnostic explanation (LIME). In this case, the authors propose a concrete implementation of local surrogate models trained to approximate the predictions of the underlying black box model. Instead of training a global surrogate model, LIME focuses on training local surrogate models to explain individual predictions. Local surrogate models have not been analyzed in this work.

6.2 Partial dependent plots applied to RF using RFE

In this section, we insert some examples of the partial dependent plot obtained in the binary classification case for the variable creatinine and basal creatinine which are in the top five most

important variables for both binary and severe AKI classification cases.

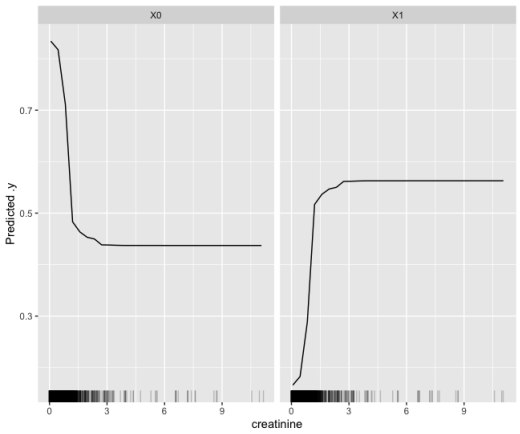

Figure 6.3: PDP of creatinine variable (binary AKI case)

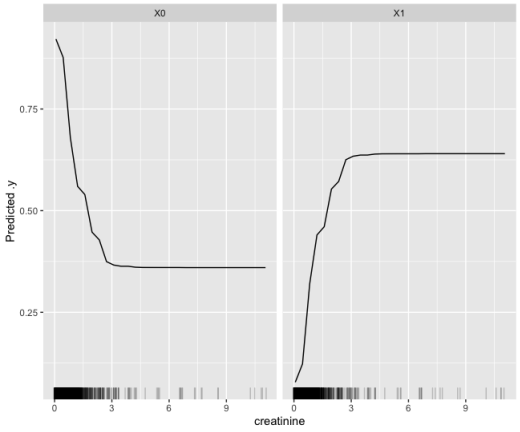

Figure 6.4: PDP of creatinine variable (severe AKI case)

From Figures 6.3 6.4 we observe that as the creatinine value increases, the possibility of being classified as class 1 increases. This nonlinear trend agrees with what was obtained using the GAM model.

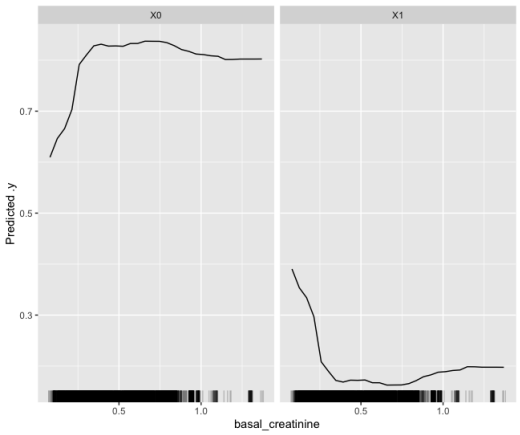

Figure 6.5: PDP of basal creatinine variable (binary AKI case)

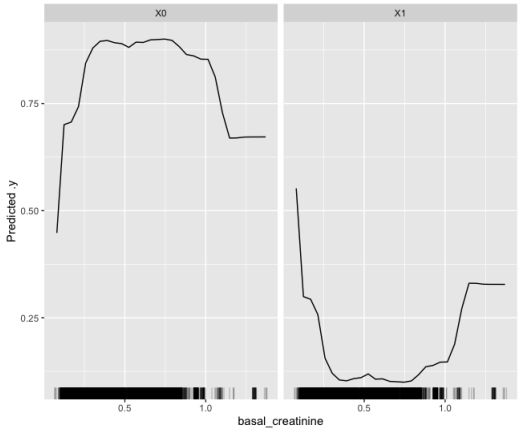

Figure 6.6: PDP of basal creatinine variable (severe AKI case)

Some reflections can also be made in the case of the basal creatinine variable. As shown in Figures 6.5 and 6.6, we have a nonlinear trend; moreover, we observe that, if the analysis is limited to the first part of the graph where the amount of data is more consistent, as the basal creatinine value increases, the possibility of being classified as class 0 increases.

A similar nonlinear trend is observed also in the multiclass case where we have a plot for each of the four classes. In Figures 6.9 and 6.10 we show the results of mode and max AKI case for the creatinine variable. Instead in Figures 6.9 and 6.10 we insert the results of basal creatinine variable. Also in this case it is possible to observe some similarities with what was observed in the case of GAMs

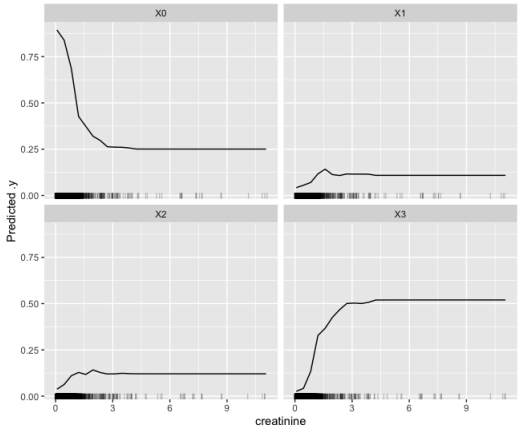

Figure 6.7: PDP of creatinine variable (mode AKI case)

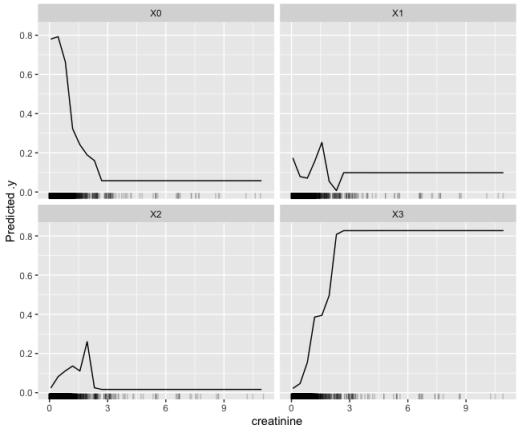

Figure 6.8: PDP of creatinine variable (max AKI case)

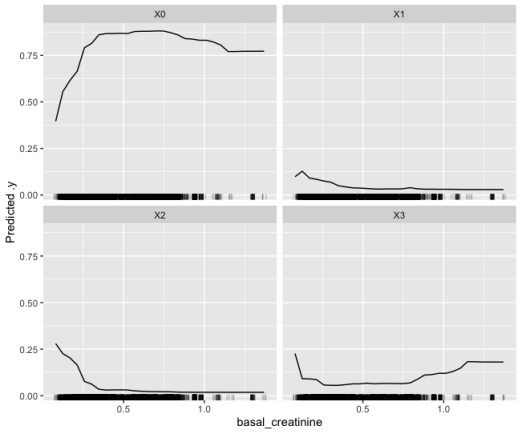

Figure 6.9: PDP of basal creatinine variable (mode AKI case)

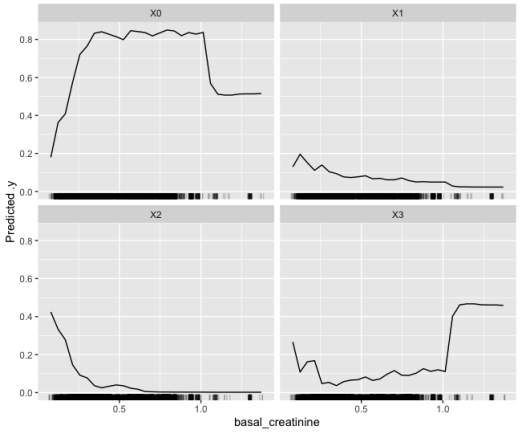

Figure 6.10: PDP of basal creatinine variable (severe AKI case)

6.3 Decision tree surrogate applied to RF using RFE

In this paragraph, we illustrate the results obtained using the decision tree surrogate approach applied to:

- binary case;

- multiclass case.

In order to maximize the prediction results we apply the DTS to the training dataset and evaluate the performance on the test set varying the maximum depth of the tree. A greater depth obviously corresponds to a more complex structure of the tree to the detriment of interpretation. For the binary class we use the AUC-ROC; starting from Figures 6.11 and 6.12 we notice that the best results are obtained with a max depth of 10.

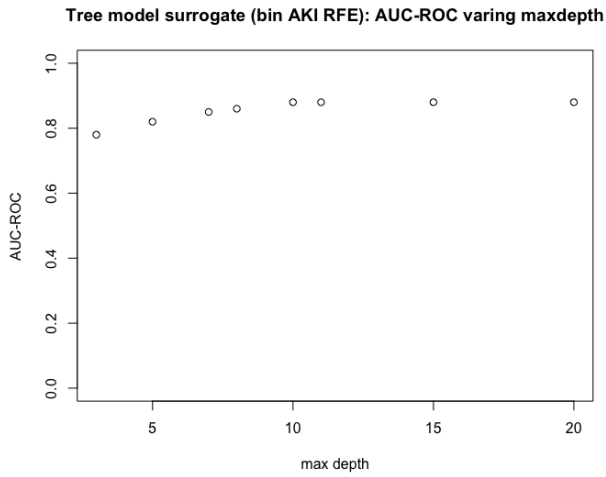

Figure 6.11: Tree model surrogate: AUC-ROC vs max depth (binary AKI case)

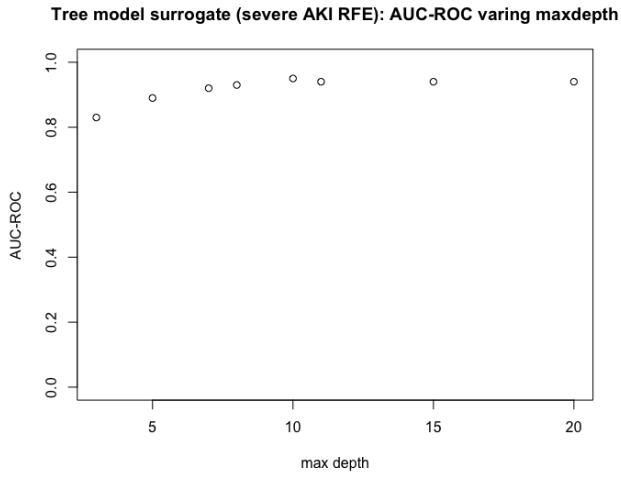

Figure 6.12: Tree model surrogate: AUC-ROC vs max depth (severe AKI case)

For the multiclass case we use as measures of performance accuracy and kappa; The values of these measures computed varying the max depth are shown in Figures 6.13 and 6.14.

We notice that the best results are obtained using:

- mode AKI case: max depth=11
- max AKI case: max depth=13

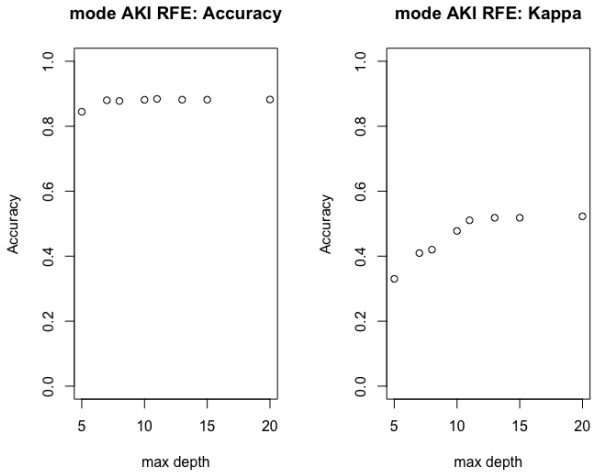

Figure 6.13: Tree model surrogate: accuracy and kappa vs max depth (mode AKI case)

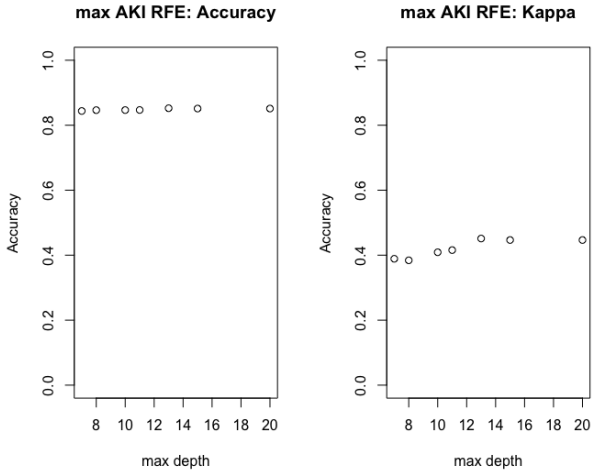

Figure 6.14: Tree model surrogate: accuracy and kappa vs max depth (max AKI case)

As for classification performance, we obtain:

- Binary AKI: AUC-ROC: 0.88
- Severe AKI: AUC-ROC: 0.95
- Mode AKI: Accuracy 0.88, Kappa 0.51
- Max AKI: Accuracy 0.85, kappa 0.45

These results are lower than those obtained with the random forest classification method.

For this reason, the Decision Tree Surrogate method, with the possible exception of the severe AKI case, cannot be considered an alternative to RFs but rather a tool that supports, together with the PDPs, the interpretation of the results under discussion with the doctors.

Chapter 7

Final discussion

This dissertation represents one of the first attempts to investigate and exploit the enormous information provided by the EMRs acquired in the PCICU of "Ospedale Pediatrico Bambino Gesù".

In particular, following the literature and the doctors' instructions, we focused on the possibility of predicting the AKI stage during the patient's stay in the intensive care unit.

In this section, we try to summarize what has been achieved by highlighting the aspects which, to the best of our knowledge, are significant and innovative compared to other studies.

Although the PCICU dataset has unique characteristics and the studies in the literature often apply to different types of patients and therefore are not directly comparable, we believe it is possible to bring out these aspects.

We try to point out these by retracing the main phases of the work presented in the thesis.

In the first phase, we approach the large size of information provided by the EMRs. This phase took a long time because a long preliminary work of knowledge and exploration of the database was necessary to identify the variables to extract. In this phase, the discussion with the doctors and with the IT staff of the company that provides the IT service of EMR was fundamental. Unfortunately, the pandemic period has made this part more complicated from a logistical point of view.

Characteristics of this phase were the choices to make a forecast over time of the stage of AKI which was not only binary but also multiclass. This multiclass prediction of max AKI and mode AKI in a given interval Δt is little explored in the literature but may offer complementary information to the binary one to support physicians' decisions.

Another important decision was to consider also the dosage of some therapies among the variables. In most of the studies we analyzed, therapies are generally not present or are present only in a qualitative way (for example presence or absence of some typology of drugs such as vasoactive ones). The extraction of the dose of therapy administered in the time interval Δt involved a further phase of comparison with doctors and IT staff.

The second phase focused on predicting the state of AKI using the RF classification. As far as we know, this is the first time that this kind of result is obtained in a PCICU. This attempt involves specific variables (e.g. PIM3, clamping duration) characteristics of this particular type of patient that differs from adult patients admitted to the intensive care unit and it has been found sometime as an important predictor. Moreover, as far as we know, our work is one of the first studies applied to the ICU dataset provided by the EMRs in Italy.

This result was by no means obvious given that we were starting from a dataset that had never

been exported with this detail with all the difficulties deriving from the interaction with the doctors and the IT staff of the company that manages the database.

The performance reported in terms of AUC-ROC and accuracy for respectively binary and multiclass classification is very good and comparable, if not superior, with those present in the literature although those studies applied to different types of patients. The RFE algorithm applied to the dataset has allowed the possibility restrict the attention to a smaller number set of variables offering physicians further possible insight and cues. Doctors could consider, also, making mandatory the acquisition of some of these sets of variables at every established interval completely avoiding the problem of possible missing or infrequent data for these measurements.

Also, the list of the most important variables was very interesting because it highlighted:

- the presence at the top of the list of some variables expected by doctors such as creatinine, basal creatinine, and ldh. This fact could be considered a further indirect confirmation of the results;
- the presence in the top ten list of the important variables of quantities specific to the PCICU case as PIM3, CEC duration shows the necessity of a specific model developed for pediatric cardiac patients;
- the presence in some cases of some therapies (adrenaline, ethacrynic acid) confirms the added value of choosing to use these quantities.

In the third phase, we move on to an analysis that relies on interpretable statical models: GAMs and BNs. Although the performances of these models are lower than those obtained with RF, they are nonetheless comparable to those presented in published studies and cited in the review made in the first chapter. The results obtained with the GAMs confirm the importance of some variables for the prediction of AKI as, for example, in the case of creatinine, basal creatinine, diuresis, aPPT_s, and moreover show a graphical trend consistent with a possible medical explanation. In the case of BN we obtain performance, especially in the severe AKI case, quite close to those of the RF with the benefit of a DAG that, taking into account the TAN constraints, shows a reasonable connection between many of the variables from a medical point of view. Further explanatory observations can be provided for medical reflection by the use of partial dependence plots as indicated in the chapter dedicated to the reflection on global methods related to the interpretable machine learning approach. The example of PDP of creatinine and basal creatinine, although they contain different information than the plots in the case of GAMs, are compatible with a possible medical explanation.

Unfortunately, the COVID pandemic has made more difficult interaction with hospital facilities. As it is well known, medical structures, especially a structure of excellence such as the "Bambino Gesù" hospital, have had to face unforeseen difficulties by limiting the time availability of doctors in projects such as this.

Based on the considerations made above, despite the difficulties derived from working on a new type of data and those deriving from the pandemic, the results confirm the possibility of making an accurate prediction of the AKI.

In fact, having obtained with different methods the result of predicting the future stage of the AKI with a set of variables that proved to be quite stable with the different methods is, in our opinion, a further confirmation of this possibility.

Moreover, we believe it is important to make a further observation: all the developed models are able to be presented to doctors using a web interface to complement or, in perspective, to be integrated into the EMR of PCICU. This tool would allow the doctors to predict the patient's stage of AKI and evaluate how to intervene if necessary.

Of course, to proceed with this, it would be necessary for the future to work, always retrospectively, on the export of the new data that has been acquired in PCICU.

Following this, a prospective study could be considered. Moreover, as shown in the literature, with the same EMR data it could be possible to analyze other medical problems.

My hope is that this study can be continued by following this path of collaboration between hospitals and universities that started during my Ph.D. This hope is confirmed by the fact that the need for technology transfer is increasingly widespread and our study could be considered an example of technology transfer in an intensive care unit, conveying in our case the methodology developed in the statistical and machine learning field to the data acquired using the EMRs.

Bibliography

Gérard Biau and Erwan Scornet. A random forest guided tour. *TEST*, 25, 11 2015. doi: 10.1007/s11749-016-0481-7.

Concha Bielza and Pedro Larrañaga. Discrete bayesian network classifiers: A survey. *ACM Comput. Surv.*, 47(1), jul 2014. ISSN 0360-0300. doi: 10.1145/2576868. URL https://doi.org/10.1145/2576868.

Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 10 2001. doi: 10.1023/A:1010950718922.

Laura Brown, Ioannis Tsamardinos, and Constantin Aliferis. A novel algorithm for scalable and accurate bayesian network learning. *Studies in health technology and informatics*, 107:711–5, 02 2004. doi: 10.3233/978-1-60750-949-3-711.

Rung-Ching Chen, Christine Dewi, Su Huang, and Rezzy Caraka. Selecting critical features for data classification based on machine learning methods. *Journal Of Big Data*, 7:26, 07 2020. doi: 10.1186/s40537-020-00327-4.

Glenn M. Chertow, Elisabeth Burdick, Melissa Honour, Joseph V. Bonventre, and David W. Bates. Acute kidney injury, mortality, length of stay, and costs in hospitalized patients. *Journal of the American Society of Nephrology*, 16(11):3365–3370, 2005. ISSN 1046-6673. doi: 10.1681/ASN.2004090740. URL https://jasn.asnjournals.org/content/16/11/3365.

CKCN Chow and Cong Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, 1968. doi: 10.1109/TIT.1968.1054142.

Francis Collins and Lawrence Tabak. Using machine learning to identify health outcomes from electronic health record data. *Nature*, 505:612–613, 2014. doi: doi.org/10.1038/505612a.

Mark Craven and Jude Shavlik. Extracting tree-structured representations of trained networks. In D. Touretzky, M.C. Mozer, and M. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8. MIT Press, 1995. URL https://proceedings.neurips.cc/paper/1995/file/45f31d16b1058d586fc3be7207b58053-Paper.pdf.

Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2005.10.010. URL https://www.sciencedirect.com/science/article/pii/S016786550500303X. ROC Analysis in Pattern Recognition.

Marine Flechet, Fabian Guiza, Miet Schetz, Pieter Wouters, Ilse Vanhorebeek, Inge Derese, Jan Gunst, Isabel Spriet, Michael Casaer, Greet Berghe, and Geert Meyfroidt. Akipredictor, an online prognostic calculator for acute kidney injury in adult critically ill patients: development, validation and comparison to serum neutrophil gelatinase-associated lipocalin. *Intensive Care Medicine*, 43, 01 2017. doi: 10.1007/s00134-017-4678-3.

Marine Flechet, Stefano Falini, Claudia Bonetti, Fabian Guiza, Miet Schetz, Greet Berghe, and Geert Meyfroidt. Machine learning versus physicians' prediction of acute kidney injury in critically ill adults: A prospective evaluation of the akipredictor. *Critical Care*, 23, 12 2019. doi: 10.1186/s13054-019-2563-x.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. ISSN 00905364. URL `http://www.jstor.org/stable/2699986`.

Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29:131–163, 11 1997. doi: 10.1023/A:1007465528199.

Joana Gameiro, Tiago Branco, and José Lopes. Artificial intelligence in acute kidney injury risk prediction. *Journal of Clinical Medicine*, 9:678, 03 2020. doi: 10.3390/jcm9030678.

Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65, 2015. doi: 10.1080/10618600.2014.907095. URL `https://doi.org/10.1080/10618600.2014.907095`.

Margherita Grandini, Enrico Bagli, and Giorgio Visani. Metrics for multi-class classification: an overview. *ArXiv*, abs/2008.05756, 2020.

Trevor Hastie and Robert Tibshirani. Generalized additive models. *Statistical Science*, 1(3):297–310, 1986. ISSN 08834237. URL `http://www.jstor.org/stable/2245459`.

Trevor Hastie and Robert Tibshirani. Generalized additive models: Some applications. *Journal of the American Statistical Association*, 82(398):371–386, 1987. ISSN 01621459. URL `http://www.jstor.org/stable/2289439`.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

Trevor J Hastie. Generalized additive models. In *Statistical models in S*, pages 249–307. Routledge, 2017.

Luke Hodgson, Nicholas Selby, Tao-Min Huang, and Lui Forni. The role of risk prediction models in prevention and management of aki. *Seminars in Nephrology*, 39:421–430, 09 2019. doi: 10.1016/j.semnephrol.2019.06.002.

Chao-Yuan Huang, Fabian Güiza Grandas, Marine Flechet, and Meyfroidt Geert. Clinical prediction models for acute kidney injury. modelos de predição clínica para lesão renal aguda na unidade de terapia intensiva: uma revisão sistemática. *Rev Bras Ter Intensiva*, 32(1):123–132., 03 2020. doi: doi:10.5935/0103-507x.20200018.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013. URL `https://faculty.marshall.usc.edu/gareth-james/ISL/`.

Christopher R John. *MLeval: Machine Learning Model Evaluation*, 2020. URL `https://CRAN.R-project.org/package=MLeval`. R package version 0.3.

Alistair E.W. Johnson, Pollard, Shen Tom J., and Lu et al. Mimic-iii, a freely accessible critical care database. *Sci Data*, 3, 2016. doi: doi.org/10.1038/sdata.2016.35.

Ahmad Kaddourah, Rajit K. Basu, Sean M. Bagshaw, and Stuart L. Goldstein. Epidemiology of acute kidney injury in critically ill children and young adults. *New England Journal of Medicine*, 376(1):11–20, 2017. doi: 10.1056/NEJMoa1611391. URL `https://doi.org/10.1056/NEJMoa1611391`. PMID: 27959707.

Arif Khwaja. Kdigo clinical practice guidelines for acute kidney injury. *Nephron Clin Pract*, pages 179–184, 2012.

Max Kuhn. The caret package. *Journal of Statistical Software*, 28, 01 2012.

Max Kuhn and Kjell Johnson. *Feature Engineering and Selection: A Practical Approach for Predictive Models (1st ed.)*. Chapman and Hall/CRC, New York, 2019. doi: https://doi.org/10.1201/9781315108230.

Vincent Labatut and Hocine Cherifi. Evaluation of performance measures for classifiers comparison. *Ubiquitous Computing and Communication Journal*, 6:21–34, 11 2011.

Vincenzo Lagani, Giorgos Athineou, Alessio Farcomeni, Michail Tsagris, and Ioannis Tsamardinos. Feature selection with the r package mxm: Discovering statistically equivalent feature subsets. *Journal of Statistical Software*, 80(7):1–25, 2017. doi: 10.18637/jss.v080.i07. URL `https://www.jstatsoft.org/index.php/jss/article/view/v080i07`.

Guiyu Lei, Guyan Wang, Congya Zhang, Yimeng Chen, and Xiying Yang. Using machine learning to predict acute kidney injury after aortic arch surgery. *Journal of Cardiothoracic and Vascular Anesthesia*, 34(12):3321–3328, 2020. ISSN 1053-0770. doi: https://doi.org/10.1053/j.jvca.2020.06.007. URL `https://www.sciencedirect.com/science/article/pii/S1053077020305139`.

Gregory Makoul, Curry, Raymond H., and Paul C. Tang. The use of electronic medical records: communication patterns in outpatient encounters. *Journal of the American Medical Informatics Association: JAMIA*, 8(6):610–615, 2001.

Kuhn Max. Building predictive models in r using the caret package. *Journal of Statistical Software*, 28, 11 2008. doi: 10.18637/jss.v028.i05.

Mary McHugh. Interrater reliability: The kappa statistic. *Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB*, 22:276–82, 10 2012. doi: 10.11613/BM.2012.031.

Patricia Mcmullen, William Howie, Nayna Philipsen, Virletta Bryant, Patricia Setlow, Mona Calhoun, and Zakevia Green. Electronic medical records and electronic health records: Overview

for nurse practitioners. *The Journal for Nurse Practitioners*, 10:660–665, 10 2014. doi: 10.1016/j.nurpra.2014.07.013.

Mihaljevi'c, Bojan, Bielza, Concha, Larranaga, and Pedro. bnclassify: Learning bayesian network classifiers. *The R Journal*, 10(2):455–468, 2018. URL `https://doi.org/10.32614/RJ-2018-073`.

Christoph Molnar. *Interpretable Machine Learning*. 2 edition, 2022. URL `https://christophm.github.io/interpretable-ml-book`.

Christoph Molnar, Bernd Bischl, and Giuseppe Casalicchio. iml: An r package for interpretable machine learning. *JOSS*, 3(26):786, 2018. doi: 10.21105/joss.00786. URL `https://joss.theoj.org/papers/10.21105/joss.00786`.

Mohammad Morid, Olivia Sheng, Guilherme Del Fiol, Julio Facelli, Bruce Bray, and Samir Abdelrahman. Temporal pattern detection to predict adverse events in critical care: A case study with acute kidney injury (preprint). *JMIR Medical Informatics*, 8, 04 2019. doi: 10.2196/14272.

Radhakrishnan Nagarajan, Marco Scutari, and Sophie Lebre. *Bayesian Networks in R: with Applications in Systems Biology*. 01 2013. ISBN 978-1-4614-6445-7. doi: 10.1007/978-1-4614-6446-4.

Jasmina Novakovic, Alempije Veljovic, Sinisa Ilic, Željko M. Papic, and Tomovic Milica. Evaluation of classification models in machine learning. *Theory and Applications of Mathematics & Computer Science*, 7:39–46, 2017.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California, June 2016a. Association for Computational Linguistics. doi: 10.18653/v1/N16-3020. URL `https://aclanthology.org/N16-3020`.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning, 2016b. URL `https://arxiv.org/abs/1606.05386`.

Lala Septem Riza, Andrzej Janusz, Christoph Bergmeir, Chris Cornelis, Francisco Herrera, Dominik Śle¸zak, and José Manuel Benítez. Implementing algorithms of rough set theory and fuzzy rough set theory in the r package "roughsets". *Information Sciences*, 287:68–89, 2014. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2014.07.029. URL `https://www.sciencedirect.com/science/article/pii/S0020025514007294`.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 12 1976. ISSN 0006-3444. doi: 10.1093/biomet/63.3.581. URL `https://doi.org/10.1093/biomet/63.3.581`.

David Ruppert, M. P. Wand, and R. J. Carroll. *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2003. doi: 10.1017/CBO9780511755453.

L Nelson Sanchez-Pinto and Robinder G Khemani. Development of a prediction model of early acute kidney injury in critically ill children using electronic health record data. *Pediatric critical care medicine : a journal of the Society of Critical Care Medicine and the World Federation of*

Pediatric Intensive and Critical Care Societies, 17(6):508—515, June 2016. ISSN 1529-7535. doi: 10.1097/pcc.0000000000000750. URL https://doi.org/10.1097/PCC.0000000000000750.

Negin Shafaf and Hamed Malek. Applications of machine learning approaches in emergency medicine; a review article. *Archives of academic emergency medicine*, 7:34, 06 2019.

Michael Simonov, Ugochukwu Ugwuowo, Erica Moreira, Yu Yamamoto, Aditya Biswas, Melissa Martin, Jeffrey Testani, and F. Perry Wilson. A simple real-time model for predicting acute kidney injury in hospitalized patients in the us: A descriptive modeling study. *PLOS Medicine*, 16(7): 1–15, 07 2019. doi: 10.1371/journal.pmed.1002861. URL https://doi.org/10.1371/journal.pmed.1002861.

Xuan Song, Xinyan Liu, Fei Liu, and Chunting Wang. Comparison of machine learning and logistic regression models in predicting acute kidney injury: A systematic review and meta-analysis. *International Journal of Medical Informatics*, 151:104484, 2021. ISSN 1386-5056. doi: https://doi.org/10.1016/j.ijmedinf.2021.104484. URL https://www.sciencedirect.com/science/article/pii/S1386505621001106.

Danielle E. Soranno, Azra Bihorac, Stuart L. Goldstein, Kianoush B. Kashani, Shina Menon, Girish N. Nadkarni, Javier A. Neyra, Neesh I. Pannu, Karandeep Singh, Jorge Cerda, and Jay L. Koyner. Artificial intelligence for aki!now: Let's not await plato's utopian republic. *Kidney360*, 3(2):376–381, 2022. doi: 10.34067/KID.0003472021. URL https://kidney360.asnjournals.org/content/3/2/376.

Daniel J. Stekhoven and Peter Bühlmann. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 10 2011. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr597. URL https://doi.org/10.1093/bioinformatics/btr597.

Mark Thomas, Caroline Blaine, Anne Dawnay, Mark Devonald, Saoussen Ftouh, Chris Laing, Susan Latchem, Andrew Lewington, David Milford, and Marlies Ostermann. The definition of acute kidney injury and its use in practice. *Kidney international*, 87:62–73, 10 2014. doi: 10.1038/ki.2014.328.

Nenad Tomašev, Xavier Glorot, Jack Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Anne Mottram, Clemens Meyer, Suman Ravuri, Ivan Protsyuk, Alistair Connell, Cían Hughes, Alan Karthikesalingam, Julien Cornebise, Hugh Montgomery, Geraint Rees, Chris Laing, Clifton Baker, Kelly Peterson, and Shakir Mohamed. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, 572:116, 08 2019. doi: 10.1038/s41586-019-1390-1.

Chen Hsi Tsai, Aboozar Eghdam, Nadia Davoody, Graham Wright, Stephen Flowerday, and Sabine Koch. Effects of electronic health record implementation and barriers to adoption and use: A scoping review and qualitative analysis of the content. *Life*, 10:327, 12 2020. doi: 10.3390/life10120327.

Ioannis Tsamardinos, Constantin F. Aliferis, and Alexander Statnikov. Time and sample efficient discovery of markov blankets and direct causal relations. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '03*, pages 673–678,

December 2003. doi: 10.1145/956750.956838. 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '03 ; Conference date: 24-08-2003 Through 27-08-2003.

Liwei et alia Wang. Impact of diverse data sources on computational phenotyping. *Frontiers in Genetics*, 11, 2020. ISSN 1664-8021. doi: 10.3389/fgene.2020.00556. URL `https://www.frontiersin.org/article/10.3389/fgene.2020.00556`.

Laura White, Wenqing Jiang, Yicheng Ma, Kaku Armah, Jeffrey Samet, and Debbie Cheng. Tutorial in biostatistics: The use of generalized additive models to evaluate alcohol consumption as an exposure variable. *Drug and Alcohol Dependence*, 209:107944, 02 2020. doi: 10.1016/j.drugalcdep.2020.107944.

Jenna Wong, Horwitz Mara Murray, Li Zhou, and Sengwee Toh. Using machine learning to identify health outcomes from electronic health record data. *Current epidemiology reports*, 5(4):331–342, 2018.

Simon N. Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):3–36, 2011. doi: https://doi.org/10.1111/j.1467-9868.2010.00749.x. URL `https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2010.00749.x`.

Simon N Wood. *Generalized Additive Models: An Introduction with R*. CRC Press, United States, 2 edition, 2017. ISBN 978-1498728331.

Simon N. Wood, Natalya Pya, and Benjamin Säfken. Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, 111(516):1548–1563, 2016. doi: 10.1080/01621459.2016.1180986. URL `https://doi.org/10.1080/01621459.2016.1180986`.

Appendix A

Classification results using RF with all variables

A.1 Binary case

In this section, we insert the classification results in the case of the RF model with all the variables in more detail. These results are obtained using MLeval R package (see John [2020]).

Table A.1: 48h classification results via Random Forest with therapies (binary AKI case)

	48h binary AKI score all variable	CI
SENS	0.712	0.68-0.74
SPEC	0.986	0.98-0.99
MCC	0.773	-
Informedness	0.697	-
PREC	0.921	0.9-0.94
NPV	0.936	0.93-0.94
FPR	0.014	-
F1	0.803	-
TP	523	-
FP	45	-
TN	3075	-
FN	212	-
AUC-ROC	0.93	0.92-0.94
AUC-PR	0.87	-
AUC-PRG	0.6	-

Table A.2: 48h classification results via Random Forest with therapies (severe AKI case)

	48h severe AKI score all variable with therapies	CI
SENS	0.74	0.7-0.78
SPEC	0.994	0.99-1
MCC	0.818	-
Informedness	0.734	-
PREC	0.946	0.92-0.96
NPV	0.965	0.96-0.97
FPR	0.006	-
F1	0.83	-
TP	350	-
FP	20	-
TN	3362	-
FN	123	-
AUC-ROC	0.99	0.98-1
AUC-PR	0.93	-
AUC-PRG	0.8	-

A.2 Multiclass case

In this section, we insert the classification results in the multiclass case in more detail.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1    2    3
         0 3084  165   52   36
         1   11   88    5    7
         2    7    6  123    4
         3   18    2    6  239

Overall Statistics

               Accuracy : 0.9172
                 95% CI : (0.9081, 0.9257)
    No Information Rate : 0.8098
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7144

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: 0 Class: 1 Class: 2 Class: 3
Sensitivity            0.9885  0.33716  0.66129  0.83566
Specificity            0.6548  0.99360  0.99536  0.99271
Pos Pred Value         0.9242  0.79279  0.87857  0.90189
Neg Pred Value         0.9302  0.95377  0.98303  0.98690
Prevalence             0.8098  0.06774  0.04827  0.07423
Detection Rate         0.8004  0.02284  0.03192  0.06203
Detection Prevalence   0.8661  0.02881  0.03634  0.06878
Balanced Accuracy      0.8217  0.66538  0.82833  0.91419
```

Table A.3: 48h confusion matrix max AKI classification results with all variables

```
          Reference
Prediction    0    1    2    3
         0 3232   83   26   43
         1    6   76    4    0
         2    2    4  136    4
         3   15    4    6  213

Overall Statistics

               Accuracy : 0.9489
                 95% CI : (0.9415, 0.9556)
    No Information Rate : 0.8446
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7968

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: 0 Class: 1 Class: 2 Class: 3
Sensitivity            0.9929  0.45509  0.79070  0.81923
Specificity            0.7462  0.99729  0.99728  0.99304
Pos Pred Value         0.9551  0.88372  0.93151  0.89496
Neg Pred Value         0.9511  0.97585  0.99029  0.98700
Prevalence             0.8446  0.04333  0.04463  0.06746
Detection Rate         0.8386  0.01972  0.03529  0.05527
Detection Prevalence   0.8780  0.02231  0.03788  0.06175
Balanced Accuracy      0.8696  0.72619  0.89399  0.90614
```

Table A.4: 48h confusion matrix mode AKI classification results with all variables

Appendix B

Classification results using a subset of variables via RFE

B.1 Binary case

In this section, we insert the classification results obtained in the binary case using a subset of variables selected using RFE in more detail.

Table B.1: 48h classification results via Random Forest using a subset of variables selected with RFE method: binary AKI case

	48h binary AKI score using RFE	CI
SENS	0.751	0.72-0.78
SPEC	0.98	0.97-0.98
MCC	0.786	-
Informedness	0.731	-
PREC	0.9	0.87-0.92
NPV	0.944	0.94-0.95
FPR	0.02	-
F1	0.819	-
TP	552	-
FP	61	-
TN	3059	-
FN	183	-
AUC-ROC	0.95	0.94-0.96
AUC-PR	0.88	-
AUC-PRG	0.65	-

Table B.2: 48h classification results via Random Forest with RFE (severe AKI case

	48h severe AKI score 10 variables selected using RFE method	CI
SENS	0.89	0.86-0.92
SPEC	0.99	0.99-1
MCC	0.894	NA
Informedness	0.88	-
PREC	0.923	0.9-0.94
NPV	0.985	0.98-0.99
FPR	0.01	-
F1	0.906	-
TP	421	-
FP	35	-
TN	3347	-
FN	52	-
AUC-ROC	0.98	0.97-0.99
AUC-PR	0.92	-
AUC-PRG	0.85	-

B.2 Multiclass case

In this section, we insert the classification results obtained using RFE for the multiclass case in more detail.

```
          Reference
Prediction    0    1    2    3
         0 3215   59   20   20
         1   15  108    1    2
         2    6    0  149   10
         3   19    0    2  228

Overall Statistics

               Accuracy : 0.96
                 95% CI : (0.9534, 0.966)
    No Information Rate : 0.8446
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8498

 Mcnemar's Test P-Value : 1.79e-07

Statistics by Class:

                     Class: 0 Class: 1 Class: 2 Class: 3
Sensitivity            0.9877  0.64671  0.86628  0.87692
Specificity            0.8347  0.99512  0.99565  0.99416
Pos Pred Value         0.9701  0.85714  0.90303  0.91566
Neg Pred Value         0.9259  0.98417  0.99377  0.99112
Prevalence             0.8446  0.04333  0.04463  0.06746
Detection Rate         0.8342  0.02802  0.03866  0.05916
Detection Prevalence   0.8599  0.03269  0.04281  0.06461
Balanced Accuracy      0.9112  0.82091  0.93097  0.93554
```

Table B.3: 48h confusion matrix mode AKI classification results obtained using a subset of variables selected via RFE

```
          Reference
Prediction    0    1    2    3
         0 3043  121   29   31
         1   50  134    6    5
         2   12    6  148    7
         3   15    0    3  243

Overall Statistics

               Accuracy : 0.926
                 95% CI : (0.9173, 0.9341)
    No Information Rate : 0.8098
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7628

 Mcnemar's Test P-Value : 8.587e-09

Statistics by Class:

                     Class: 0 Class: 1 Class: 2 Class: 3
Sensitivity            0.9753  0.51341  0.79570  0.84965
Specificity            0.7531  0.98302  0.99318  0.99495
Pos Pred Value         0.9439  0.68718  0.85549  0.93103
Neg Pred Value         0.8776  0.96528  0.98967  0.98803
Prevalence             0.8098  0.06774  0.04827  0.07423
Detection Rate         0.7898  0.03478  0.03841  0.06307
Detection Prevalence   0.8368  0.05061  0.04490  0.06774
Balanced Accuracy      0.8642  0.74821  0.89444  0.92230
```

Table B.4: 48h confusion matrix max AKI classification results obtained using a subset of variables selected via RFE

Appendix C

Classification results using GAMs

C.1 Binary case

In this section, we insert the classification results obtained using GAMs in the binary case in more detail.

Table C.1: 48h classification results of GAM model with all variables (binary AKI case)

	48h binary AKI score all variable	CI
SENS	0.510	0.47-0.55
SPEC	0.965	0.96-0.97
MCC	0.563	-
Informedness	0.475	-
PREC	0.775	0.74-0.81
NPV	0.893	0.88-0.9
FPR	0.035	-
F1	0.615	-
TP	375	-
FP	109	-
TN	3011	-
FN	360	-
AUC-ROC	0.870	0.85-0.89
AUC-PR	0.700	-
AUC-PRG	0.300	-

Table C.2: 48h classification results of GAM model with all variables (severe AKI case)

	48h binary AKI score all variable	CI
SENS	0.573	0.53-0.62
SPEC	0.979	0.97-0.98
MCC	0.637	-
Informedness	0.552	-
PREC	0.792	0.75-0.83
NPV	0.942	0.93-0.95
FPR	0.021	-
F1	0.665	-
TP	271	-
FP	71	-
TN	3311	-
FN	202	-
AUC-ROC	0.940	0.92-0.96
AUC-PR	0.760	-
AUC-PRG	0.410	-

C.2 Multiclass case

In this section, we insert the classification results obtained using GAMs in the multiclass case in more detail.

```
pc_mode    0    1    2    3
      0 3189  120   77   74
      1   12   22    1    4
      2   27   10   86    6
      3   30    4   11  182

Overall Statistics

               Accuracy : 0.9025
                 95% CI : (0.8927, 0.9117)
    No Information Rate : 0.8451
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5858

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: 0 Class: 1 Class: 2 Class: 3
Sensitivity            0.9788 0.141026  0.49143  0.68421
Specificity            0.5461 0.995404  0.98832  0.98746
Pos Pred Value         0.9217 0.564103  0.66667  0.80176
Neg Pred Value         0.8253 0.964885  0.97611  0.97685
Prevalence             0.8451 0.040467  0.04540  0.06900
Detection Rate         0.8272 0.005707  0.02231  0.04721
Detection Prevalence   0.8975 0.010117  0.03346  0.05888
Balanced Accuracy      0.7624 0.568215  0.73987  0.83584
```

Table C.3: 48h confusion matrix mode AKI classification results obtained using GAM

```
Confusion Matrix and Statistics


pc_max    0    1    2    3
     0 3056  222   82   80
     1   14   19    1    2
     2   22   11   93    9
     3   30    2    9  203

Overall Statistics

               Accuracy : 0.8744
                 95% CI : (0.8636, 0.8848)
    No Information Rate : 0.8099
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5353

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: 0 Class: 1 Class: 2 Class: 3
Sensitivity            0.9789 0.074803  0.50270  0.69048
Specificity            0.4761 0.995279  0.98856  0.98849
Pos Pred Value         0.8884 0.527778  0.68889  0.83197
Neg Pred Value         0.8410 0.938466  0.97527  0.97480
Prevalence             0.8099 0.065888  0.04799  0.07626
Detection Rate         0.7927 0.004929  0.02412  0.05266
Detection Prevalence   0.8923 0.009339  0.03502  0.06329
Balanced Accuracy      0.7275 0.535041  0.74563  0.83948
```

Table C.4: 48h confusion matrix max AKI classification results obtained using GAM

Appendix D

Classification results using BN classifiers

D.1 binary case

In this section, we insert the classification results obtained using BN with all the variables in the binary case in more detail.

	48h binary AKI score with BN	CI
SENS	0.721	0.69-0.75
SPEC	0.896	0.89-0.91
MCC	0.584	NA
Informedness	0.618	NA
PREC	0.621	0.59-0.65
NPV	0.932	0.92-0.94
FPR	0.104	NA
F1	0.668	NA
TP	530	NA
FP	323	NA
TN	2797	NA
FN	205	NA
AUC-ROC	0.900	0.88-0.92
AUC-PR	0.740	NA
AUC-PRG	0.370	NA

Table D.1: AUC ROC binary AKI case

	48h severe AKI score with BN	CI
SENS	0.822	0.79-0.85
SPEC	0.939	0.93-0.95
MCC	0.691	NA
Informedness	0.761	NA
PREC	0.653	0.61-0.69
NPV	0.974	0.97-0.98
FPR	0.061	NA
F1	0.728	NA
TP	389	NA
FP	207	NA
TN	3175	NA
FN	84	NA
AUC-ROC	0.970	0.96-0.98
AUC-PR	0.820	NA
AUC-PRG	0.530	NA

Table D.2: AUC ROC severe AKI case

D.2 Multiclass case

In this section, we insert the classification results obtained using BN in the multiclass case in more detail.

```
pc_mode    0    1    2    3
      0 3097   62   26   20
      1   50   98    7    1
      2   54    2  132   10
      3   54    5    7  229

Overall Statistics

               Accuracy : 0.9227
                 95% CI : (0.9138, 0.9309)
    No Information Rate : 0.8446
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7319

 Mcnemar's Test P-Value : 1.208e-05

Statistics by Class:

                     Class: 0 Class: 1 Class: 2 Class: 3
Sensitivity            0.9515  0.58683  0.76744  0.88077
Specificity            0.8197  0.98427  0.98207  0.98164
Pos Pred Value         0.9663  0.62821  0.66667  0.77627
Neg Pred Value         0.7565  0.98134  0.98906  0.99129
Prevalence             0.8446  0.04333  0.04463  0.06746
Detection Rate         0.8036  0.02543  0.03425  0.05942
Detection Prevalence   0.8316  0.04048  0.05138  0.07654
Balanced Accuracy      0.8856  0.78555  0.87476  0.93120
```

Table D.3: 48h confusion matrix mode AKI classification results obtained using BN with all the variables.

```
pc_max    0    1    2    3
     0 2891  123   41   25
     1  105  118   10    5
     2   74   13  133   19
     3   50    7    2  237

Overall Statistics

               Accuracy : 0.877
                 95% CI : (0.8662, 0.8872)
    No Information Rate : 0.8098
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.638

 Mcnemar's Test P-Value : 7.651e-06

Statistics by Class:

                     Class: 0 Class: 1 Class: 2 Class: 3
Sensitivity            0.9266  0.45211  0.71505  0.82867
Specificity            0.7422  0.96659  0.97109  0.98346
Pos Pred Value         0.9386  0.49580  0.55649  0.80068
Neg Pred Value         0.7038  0.96044  0.98533  0.98622
Prevalence             0.8098  0.06774  0.04827  0.07423
Detection Rate         0.7503  0.03063  0.03452  0.06151
Detection Prevalence   0.7994  0.06177  0.06203  0.07682
Balanced Accuracy      0.8344  0.70935  0.84307  0.90607
```

Table D.4: 48h confusion matrix max AKI classification results obtained using BN with all the variables.

Appendix E

Classification results using BN classifier and MXM subset of variables

E.1 Binary AKI case

In this section, we insert the classification results obtained using BN and MXM subset of variables in the binary case in more detail.

	48h binary AKI score with BN and MXM subset	CI
SENS	0.729	0.7-0.76
SPEC	0.890	0.88-0.9
MCC	0.580	NA
Informedness	0.620	NA
PREC	0.610	0.58-0.64
NPV	0.933	0.92-0.94
FPR	0.110	NA
F1	0.665	NA
TP	536	NA
FP	342	NA
TN	2778	NA
FN	199	NA
AUC-ROC	0.900	0.88-0.92
AUC-PR	0.730	NA
AUC-PRG	0.340	NA

Table E.1: BN AUC-ROC bin AKI with MXM subset

	48h severe AKI score with BN and MXM subset	CI
SENS	0.799	0.76-0.83
SPEC	0.948	0.94-0.96
MCC	0.700	NA
Informedness	0.747	NA
PREC	0.684	0.64-0.72
NPV	0.971	0.96-0.98
FPR	0.052	NA
F1	0.737	NA
TP	378	NA
FP	175	NA
TN	3207	NA
FN	95	NA
AUC-ROC	0.970	0.96-0.98
AUC-PR	0.820	NA
AUC-PRG	0.550	NA

Table E.2: BN AUC-ROC severe AKI with MXM subset

E.2 Multiclass case

In this section, we insert the classification results obtained using BN and MXM subset of variables in the multiclass case in more detail.

```
pc_mode    0    1    2    3
      0 3050   58   20   18
      1   84  101    5    1
      2   65    6  142    9
      3   56    2    5  232

Overall Statistics

               Accuracy : 0.9146
                 95% CI : (0.9054, 0.9233)
    No Information Rate : 0.8446
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7161

 Mcnemar's Test P-Value : 5.488e-09

Statistics by Class:

                     Class: 0 Class: 1 Class: 2 Class: 3
Sensitivity            0.9370  0.60479  0.82558  0.89231
Specificity            0.8397  0.97559  0.97827  0.98247
Pos Pred Value         0.9695  0.52880  0.63964  0.78644
Neg Pred Value         0.7105  0.98198  0.99174  0.99213
Prevalence             0.8446  0.04333  0.04463  0.06746
Detection Rate         0.7914  0.02621  0.03684  0.06020
Detection Prevalence   0.8163  0.04956  0.05760  0.07654
Balanced Accuracy      0.8884  0.79019  0.90193  0.93739
```

Table E.3: 48h confusion matrix mode AKI classification results obtained using BN with a subset of variables selected using MXM.

```
pc_max    0    1    2    3
     0 2861  121   32   35
     1  148  124   10    3
     2   62   11  139   11
     3   49    5    5  237

Overall Statistics

               Accuracy : 0.8723
                 95% CI : (0.8614, 0.8827)
    No Information Rate : 0.8098
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6306

 Mcnemar's Test P-Value : 0.007872

Statistics by Class:

                     Class: 0 Class: 1 Class: 2 Class: 3
Sensitivity            0.9170  0.47510  0.74731  0.82867
Specificity            0.7435  0.95518  0.97709  0.98346
Pos Pred Value         0.9383  0.43509  0.62332  0.80068
Neg Pred Value         0.6779  0.96160  0.98705  0.98622
Prevalence             0.8098  0.06774  0.04827  0.07423
Detection Rate         0.7425  0.03218  0.03608  0.06151
Detection Prevalence   0.7913  0.07397  0.05788  0.07682
Balanced Accuracy      0.8303  0.71514  0.86220  0.90607
```

Table E.4: 48h confusion matrix max AKI classification results obtained using BN with a subset of variables selected using MXM.

Acknowledgements

First of all, I would like to thank all the doctors of the intensive care unit of Bambino Gesù Hospital, and in particular Tiziana Fragasso, Valeria Raggi, and Zaccaria Ricci, for their availability and listening skills.

A special thanks also to Leonida Passeri of Ascom for his support on the PCICU database.

I would also like to thank Dr. Alberto Tozzi and Dr. Caterina Rizzo, in particular, for their initial support.

On the university front, I would like to thank Prof. Marco Alfò for welcoming me into the Ph.D. program and Prof. Stefania Gubbiotti for her advice.

A special thank goes to Prof. Giovanna Jona Lasinio for agreeing to guide me in the initial phase of this thesis project, and of course to Prof. Luca Tardella, who supported me all the way and showed increasing interest in the project.

I would also like to thank Prof. Francesco Lagona for the availability he gave me to follow me in some attempts that did not eventually make it into this thesis.

I would like to thank all the "young doctoral students" with whom I had the opportunity to compare myself, and in particular: Luca Merlo, Michela Cianfriglia, Giorgia Zaccaria, and Gianmarco Caruso.

Special thanks go to Maria Luigia Menditto for the mutual support during the Ph.D. and to Lorenzo Mazza and Antonio Veredice, with whom I had the opportunity to discuss topics related to mathematics education.

Thanks also to Enrico Carta for his database crash course and Silvia Maschio for her tips on LaTeX.

Finally, I would like to thank my wife, Stefania, without whose patience and support I would not have made it this far, and my sons Mattia and Giacomo, who, without knowing it, gave me the joy and strength to carry on with this project.