Machine Learning based Classification Models for COVID-19 Patients

Francesca Maggioni[1], Daniel Faccini[1], Federico Gheza[2], Filippo Manelli[3], and Graziella Bonetti[4]

[1] Department of Management, Information and Production Engineering, University of Bergamo; Francesca.Maggioni@unibg.it, Daniel.Faccini@unibg.it

[2] Research Fellow in General Surgery and Staff Surgeon, University of Brescia; Federico.Gheza1@unibs.it

[3] Director of Emergency Unit, ASST-Bergamo Est, Italy; Filippo.Manelli@asst-bergamoest.it

[4] Director of Clinical Pathology Laboratory, ASST-Valcamonica, Esine, Italy; Graziella.Bonetti@asst-valcamonica.it

Abstract. The SARS-CoV-2 pandemic has pushed the National Health Service to extraordinary pressure, causing situations of imbalance between the request and availability of assistance. When the number of patients exceeds the available resources, doctors need to establish priorities among the patients to be treated. This paper describes novel data-driven optimization models to support doctors' decisions to solve one of the main problems encountered during the first months of the COVID-19 pandemic: predict the mortality risk for COVID-19 in order to address the most appropriate therapeutic path. The models are trained using clinical data obtained at the access to the Emergency Department of 150 SARS-CoV-2 infected patients admitted to ASST-Valcamonica (Brescia, Italy), in March 2020. To handle the uncertainty in data, we formulate robust and distributionally robust optimization models and compare their performance with other 31 different classification models from the literature, including decision trees, discriminant analysis, support vector machines, logistic regression, nearest neighbors, and naive Bayes. Numerical results show that robust formulations allow to achieve higher levels of accuracy with respect to the corresponding deterministic ones. The best prediction results are obtained with an optimized decision tree model, allowing to identify the most important factors. The tool can be used after triage to more accurately assess the severity of a COVID-19 patient's condition, allowing doctors to optimize patient accommodation by identifying those in need of intensive care and those instead of sub-intensive care.

Keywords: Decision Support · COVID-19 Application · Data Analysis and Risk Management.

1 Introduction

Coronaviruses are RNA viruses, known to cause disease in humans and animals, ranging from common cold to more severe and even deadly respiratory

infections. Strains of betacoronavirus have been identified in 2003 and 2012 as causing *Severe Acute Respiratory Syndrome* (SARS) and also the *Middle East Respiratory Syndrome* (MERS). *Coronavirus Disease 2019* (COVID-19) is an infectious disease caused by *SARS Coronavirus 2* (SARS-CoV-2) according to the International Committee on Taxonomy of Viruses [4]. SARS-CoV-2 is a new type of coronavirus and its nucleic acid sequence is different from SARS-CoV and MERS-CoV. At April 6^{th} 2020, the last statistics of the *World Health Organization* (WHO) revealed that COVID-19 had already affected over 554.550 people from almost every country worldwide, causing as many as 47.687 death, while as of May 31^{st} 2022 cases reached 532.941.055 with around 6.313.941 deaths. As of June 2022, Italy is the 9^{th} of the most deeply involved country, for the number of cases (17.421.410), deaths (166.697) and especially the mortality rate, 2.765 deaths every one million citizens [10].

The clinical manifestations of COVID-19 infection include fever, myalgia, dry cough, dyspnoea, fatigue and less frequently headache, diarrhoea, nausea, vomiting as well as anosmia and ageusia. In severe cases, COVID-19 can rapidly turn into acute respiratory distress syndrome, septic shock, bleeding, coagulation dysfunction, metabolic acidosis and death [6]. The role of laboratory medicine has always been of critical importance during viral outbreaks [7], due to its ability of identifying possible clinical predictors of progression towards severe and fatal forms of infections [1]. Some parameters which have already been shown to have an influence on COVID-19 patients outcomes are: aspartate aminotransferase, lactate dehydrogenase, C-reactive protein, neutrophils and lymphocyte counts, haemoglobin, platelets count, procalcitonin, high sensitive cardiac troponin I, urea, creatinine, cardiac biomarkers, and partial hromboplastin time [3]. The definition of the strongest predictors enables risk stratification among patients at high or low risk of mortality, allowing for improved clinical situational care.

Therefore, the aim of the present study is to analyze the common laboratory abnormalities in patients with COVID-19 employing the most recent *Machine Learning* (ML) techniques, in order to identify which are the parameters most likely to classify patients between those who are well and those who are unlikely to survive. From the ML standpoint, a great variety of algorithms have been devised to address the classification problem: *Decision Trees* (DT), *Discriminant Analysis* (D), *Logistic Regression* (LR), *Naive Bayes* (NB), *Support Vector Machines* (SVM), and *k-Nearest Neighbors* (k-NN) classifiers, etc. An underlying assumption of ML approaches is to handle noise in data only indirectly (or implicitly) at the moment of classifying. This assumption, however, is not always practical as real-world observations are often plagued by noise (*e.g.*, due to limited precision of collecting instruments, measurement mistakes in data gathering, sampling errors, etc.) and two of the main paradigms to deal with problems affected with uncertainty are given by *Robust Optimization* (RO) [2] and *Distributionally Robust Optimization* (DRO) [9].

In this paper, therefore, to predict COVID-19 mortality risk and support doctors' decisions of addressing the most appropriate patients' therapeutic path, we perform a computational comparison of literature ML classification models,

including novel robust and distributionally robust SVM formulations from [5] that explicitly handle data uncertainty in the training set.

The paper is organized as follows. Section 2 presents robust and distributionally robust optimization models for SVM; Section 3 describes data collection and reports the experimental study on prediction of COVID-19 mortality risk, while conclusions are provided in Section 4.

2 Methods

In this section we aim to build a ML model to support doctors' decisions of "stratifying patients' clinical risk", hence to predict the mortality risk of COVID-19 patients so as to guide the best diagnostic and therapeutic care path. Specifically, we will first recall the deterministic formulation from [8]. To handle uncertainty in data features due to limited precision of collecting instruments or measurement mistakes, we further consider the robust and distributionally robust counterparts proposed in [5].

2.1 Deterministic Formulation

Let $X := \{x^{(1)}, x^{(2)}, \ldots, x^{(I)}\} \subseteq \mathbb{R}^n$ and $Y := \{y^{(1)}, y^{(2)}, \ldots, y^{(J)}\} \subseteq \mathbb{R}^n$ be two sets which correspond, respectively, to surviving COVID-19 patients (class 0) and to COVID-19 patients who died within their hospital stay (class 1). For every patient n features based on clinical and laboratory data are available. The hyperplane $a^\top x = \gamma$ (with $a \in \mathbb{R}^n, \gamma \in \mathbb{R}$ later denoted with (a, γ)) that separates sets X and Y is found solving the following deterministic SVM optimization problem (see [8]):

$$\begin{array}{lll} \min\limits_{a,\gamma,z_X,z_Y} & \|a\|_1 + \nu(e^\top z_X + e^\top z_Y) & \\ \text{s.t.} & a^\top x^{(i)} \leq \gamma - 1 + z_{x^{(i)}} & i = 1, \ldots, I \\ & a^\top y^{(j)} \geq \gamma + 1 - z_{y^{(j)}} & j = 1, \ldots, J \\ & z_X, z_Y \geq 0, & \end{array} \tag{1}$$

where $\|\cdot\|_1$ denote the 1-norm, vector e has all entries equal to one, $z_X := [z_{x^{(1)}}; \ldots; z_{x^{(I)}}] \in \mathbb{R}^I_+$ and $z_Y := [z_{y^{(1)}}; \ldots; z_{y^{(J)}}] \in \mathbb{R}^J_+$ are the non-negative vectors of errors of group X and Y. Observation $x^{(i)} \in \mathbb{R}^n$ is correctly classified if $0 \leq z_{x^{(i)}} \leq 1$, misclassified otherwise. Further, $\nu \geq 0$ is a user-defined penalty parameter allowing a trade-off between the margin maximization ($\|a\|_1$) and tolerating misclassification ($e^\top z_X + e^\top z_Y$). Once the starting hyperplane (a, γ) is obtained, it is shifted in order to determine hyperplane $H_1 := (a, \gamma - 1 + \omega_1)$ such that all points corresponding to patients in X lie on one of its side, and hyperplane $H_2 := (a, \gamma + 1 - \omega_2)$ such that all points corresponding to patients in Y lie on the opposite side. Finally, through line search hyperplane H_3 is identified lying between H_1 and H_2 and minimizing the overall number of misclassified patients.

2.2 Robust Formulation

To handle uncertainty in data features, we now consider a robust counterpart of model (1) with uncertainty sets in the form of hyperrectangles and hyperellipsoids (see [5]). We assume the uncertainty of every patient data $x^{(i)} \in X \subseteq \mathbb{R}^n,\ i = 1, \dots, I$ to be represented by the uncertainty set $\mathcal{U}\big(x^{(i)}\big)$. Equivalently for every $y^{(j)},\ j = 1, \dots, J$. Then, the robust counterpart of model (1) corresponds to the following optimization model:

$$\begin{aligned} \min_{a,\gamma,z_X,z_Y} \quad & \|a\|_1 + \nu(e^\top z_X + e^\top z_Y) \\ \text{s.t.} \quad & \max_{x \in \mathcal{U}(x^{(i)})} \left[a^\top x\right] \leq \gamma - 1 + z_{x^{(i)}} \qquad i = 1, \dots, I \\ & \min_{y \in \mathcal{U}(y^{(j)})} \left[a^\top y\right] \geq \gamma + 1 - z_{y^{(j)}} \qquad j = 1, \dots, J \\ & z_X, z_Y \geq 0. \end{aligned} \tag{2}$$

Uncertainty sets are built as follows:

- Let $\zeta_{x^{(i)}} \in \mathbb{R}^n_+$ define the perturbation vector of $x^{(i)}$ and let $\rho_X \in \mathbb{R}_+$ be the global measure of uncertainty for group X. Then, the hyperrectangular uncertainty set $\mathcal{U}_{\mathcal{B}}\big(x^{(i)}\big)$ centered around $x^{(i)}$ is defined as:

$$\mathcal{U}_{\mathcal{B}}\big(x^{(i)}\big) := \left\{x \in \mathbb{R}^n \;\middle|\; x^{(i)} - \rho_X \zeta_{x^{(i)}} \leq x \leq x^{(i)} + \rho_X \zeta_{x^{(i)}}\right\} \tag{3}$$

  and equivalently for every $y^{(j)},\ j = 1, \dots, J$.
- Let $\Sigma_{x^{(i)}} \in \mathbb{R}^{n \times n}$ be the positive definite covariance matrix associated to $x^{(i)}$. Then, the ellipsoidal uncertainty set $\mathcal{U}_{\mathcal{E}}\big(x^{(i)}\big)$ centered around $x^{(i)}$ with ray $\rho_X \in \mathbb{R}_+$ is defined as:

$$\mathcal{U}_{\mathcal{E}}\big(x^{(i)}\big) := \left\{x \in \mathbb{R}^n \;\middle|\; \big(x - x^{(i)}\big)^\top \Sigma^{-1}_{x^{(i)}} \big(x - x^{(i)}\big) \leq \rho_X^2\right\} \tag{4}$$

  and equivalently for every $y^{(j)},\ j = 1, \dots, J$.

Once the starting hyperplane (a, γ) determined in (2) is obtained, it is shifted in order to determine hyperplane H_1, H_2 and H_3 as described above.

2.3 Distributionally Robust Formulation

We now consider a less conservative way to handle uncertainty in data features by considering a distributionally robust counterpart of model (1), where we treat all input patients data $x^{(i)},\ i = 1, \dots, I$ as random variables for which the exact probability distributions $\mathbb{P}^{\text{true}}_{x^{(i)}},\ i = 1, \dots, I$ are unknown (see [5]). For each $x^{(i)}$ we optimize against the worst-case expectation under all possible distributions $\mathbb{P}$ belonging to the ambiguity set $\mathcal{D}\big(x^{(i)}\big)$. Equivalently for $y^{(j)}$ and $\mathcal{D}\big(y^{(j)}\big)$. Accordingly, the distributionally robust counterpart of model (1) can

be formulated as follows:

$$\begin{array}{ll} \min\limits_{a,\gamma,z_X,z_Y} & \|a\|_1 + \nu(e^\top z_X + e^\top z_Y) \\ \text{s.t.} & \sup\limits_{\mathbb{P}\in\mathcal{D}(x^{(i)})} \mathbb{E}_{\mathbb{P}}\big[a^\top x\big] \le \gamma - 1 + z_{x^{(i)}} \qquad i = 1,\dots,I \\ & \inf\limits_{\mathbb{P}\in\mathcal{D}(y^{(j)})} \mathbb{E}_{\mathbb{P}}\big[a^\top y\big] \ge \gamma + 1 - z_{y^{(j)}} \qquad j = 1,\dots,J \\ & z_X, z_Y \ge 0. \end{array} \tag{5}$$

We consider the general class moment-based ambiguity set proposed in [9] where the support and a list of partial moments describing the uncertainty are available:

$$\mathcal{D}\big(x^{(i)}\big) := \left\{ \mathbb{P} \in \mathcal{P}_+^n \;\middle|\; \begin{array}{c} \mathbb{P}\big(x \in \mathcal{U}_{\mathcal{B}}\big(x^{(i)}\big)\big) = 1 \\ \mathbb{E}_{\mathbb{P}}\big[g_p(x)\big] \le \big(\varrho_X\big)_p \quad p = 1,\dots,n \end{array} \right\}$$

with $\mathcal{P}_+^n$ representing the set of probabilities distributions on $\mathbb{R}^n$. Specifically, the first constraint requires every realization to be constrained within its support set $\mathcal{U}_{\mathcal{B}}\big(x^{(i)}\big)$ defined as in (3), while the second group of constraints characterizes the moments information via n functions $g_p(\cdot)$, and enforces the generalized moment $\mathbb{E}_{\mathbb{P}}\big[g_p(x)\big]$ not to exceed a given threshold $(\varrho_X)_p \in \mathbb{R}_+$, $p = 1,\dots,n$. The moment function we employ in this paper is the piecewise linear formulation which can be interpreted as the first-order deviations of the uncertain parameter with respect to the nominal patient value $x^{(i)}$ along projections $f_X^{(p)} \in \mathbb{R}^n$: $g_p(x) := |f_X^{(p)\top}(x - x^{(i)})|$, $p = 1,\dots,n$. To determine projections $f_X^{(1)},\dots,f_X^{(n)}$ and thresholds $(\varrho_X)_1,\dots,(\varrho_X)_n$ we adopt the database strategy based on *Principal Component Analysis* (PCA), described in [5]. Equivalently for every $y^{(j)}$, $j = 1,\dots,J$. Model (5) is intractable due to the infinite number of probability distributions contained in every ambiguity set; therefore, in [5], a tractable deterministic reformulation is provided.

Among the proposed formulations, model (2) with hyperrectangles uncertainty sets leads to the most conservative solutions. On the other hand, hyperellipsoids uncertainty sets lead to less conservative solutions, as those situations under which all features jointly assume extreme interval values are disregarded. Finally, the distributionally robust model (5) represents the most aggressive approach among the three, since it does not only rely on support information, but rather also assumes to have knowledge about the moments distributions.

3 Experimental Study

3.1 Data Collection

From an initial group of 150 COVID-19 patients admitted to the *Emergency Room* (ER) of Valcamonica Hospital (Esine, Brescia - Lombardy, Italy) between March 5$^{\text{th}}$ and April 1$^{\text{st}}$ 2020 and diagnosed with COVID-19, two different cohorts were selected for having complete clinical and laboratory data. The first cohort included 78 COVID-19 patients (47 males, 31 females, age range $32-90$)

who survived and could hence be discharged, whilst the second group encompassed 72 patients (20 males, 52 females, age range $55-95$) who died within their hospital stay. All patients were diagnosed with COVID-19 according to current standards, *i.e.*, displaying suggestive findings at chest *Computed Tomography* (CT; the classic ground glass pattern of interstitial pneumonia for a minimum of $35\% - 40\%$ of lung parenchyma) and positive results of real-time *Reverse Transcriptase Polymerase Chain Reaction* (RT-PCR) for SARS-CoV-2. Real-time RT-PCR was used for direct virus identification in nasopharyngeal swabs and was performed in reference laboratories from Lombardy network of COVID-19 diagnostics. For every patient, clinical history, signs, symptoms and results of laboratory investigations were collected from the local *Laboratory Information System* (LIS), as well as from medical records, using a standard form adopted for reporting data on infectious diseases to the Italian Ministry of Health. All clinical and laboratory data here described have been recorded upon hospital admission, and the quality of collection results was validated with *Internal Quality Control* (IQC) procedures and participation to the *External Quality Assessment Scheme* (EQAS) of Lombardy region, Italy. Clinical and laboratory information was gathered during clinical workout and the study was preliminary approved by the Ethical Committee of Brescia (certificate no. NP 4036). The study was carried out in accordance with the revised declaration of Helsinki and with the term of local legislation.

3.2 Numerical Investigation

Models of Section 2 have been trained using data of patients described above. First we perform a selection of input features, then we validate the models. The computations have been performed on a 64-bit machine with 8 GB of RAM, a 1.8 GHz Intel i7 processor, and numerical results are obtained under Matlab R2022a environment using MOSEK solver (version 8.1.0.72).

Feature Selection The attributes considered are comorbidities, vital and laboratory parameters. Specifically, the following 26 predictors have been initially examined: test COVID-19 (positive, negative, doubted), sex (male, female), age (range $32 - 95$), presence of chronic diseases (yes, no), presence of neoplasia (yes, no), diabetes (yes, no), presence of cardiovascular diseases (yes, no), immunodeficiency (yes, no), presence of respiratory diseases (yes, no), presence of kidney diseases (yes, no), presence of metabolic diseases (yes, no), body mass index BMI in the range of $30 - 40$ (yes, no), body mass index BMI greater than 40 (yes, no), white blood cell count WBC (range $2{,}18 - 25{,}53$), platelet count PLT (range $64 - 521$), neutrophils (range $1{,}31 - 21{,}43$), lymphocytes (range $0{,}27 - 2{,}77$), D-dimero (range $270 - 20.000$), aspartate aminotransferase AST (range $9 - 464$), lactate dehydrogenase LDH (range $117 - 1.161$), creatine kinase CK (range $7 - 4.038$), C-reactive protein PCR (range $8{,}3 - 372{,}6$), high sensitivity cardiac troponin I cTnI (range $5 - 5.124$), ferritin (range $127 - 8.413$), WBC/lymphocytes % (range $1{,}78 - 46{,}95$) and emogas P/F (range $36 - 687{,}14$).

Since too many input attributes can cause overfitting and consequent poor performance of the ML algorithms, a χ-square test has first been performed,

which allows reducing the number of input features by identifying whether each predictor is independent of the response variable (survived/deceased). According to the χ-square test, the following 15 attributes are the most important: cTnI, LDH, P/F, WBC/lymphocytes %, age, D-dimero, presence of chronic diseases, AST, neutrophils, diabetes, test COVID-19, PCR, presence of kidney diseases, CK and presence of respiratory diseases.

Models Validation Models have been trained using a 50-folds cross-validation scheme: the data sample has been split into 50 sub-groups $i = 1, \dots, 50$ each containing exactly 3 observations; the first subset is then used to validate the model that has been trained using the remaining groups $i = 2, \dots, 50$. The process is repeated 50 times so that each subset i is used exactly once for validation. The goodness of the models is evaluated by measuring the indicators associated with group i:

$$\mathrm{A}_i := \frac{\mathrm{TP}_i + \mathrm{TN}_i}{\mathrm{TP}_i + \mathrm{TN}_i + \mathrm{FP}_i + \mathrm{FN}_i}, \quad \mathrm{R}_i := \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FN}_i}, \quad \mathrm{P}_i := \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FP}_i},$$

where TP_i stands for true positive, TN_i for true negative, FP_i for false positive and FN_i for false negative of group i. Finally, scores have been averaged as:

$$\text{Accuracy} := \frac{\sum_{i=1}^{50} \mathrm{A}_i}{50}, \quad \text{Recall} := \frac{\sum_{i=1}^{50} \mathrm{R}_i}{50}, \quad \text{Precision} := \frac{\sum_{i=1}^{50} \mathrm{P}_i}{50}.$$

For each formulation of Section 2, we report in Table 1 the indicator scores for increasing levels of robustness ρ_X, ρ_Y. For all formulations the user-defined penalty parameter $\nu = 1{,}5\mathrm{e}^{-2}$. For robust models, perturbation vectors $\zeta_{x^{(i)}}, \zeta_{y^{(j)}}$ and covariance matrices $\Sigma_{x^{(i)}}, \Sigma_{y^{(j)}}$ are set to be proportional to group X and Y standard deviations, while for the distributionally robust formulation moment thresholds are proportional to a scale factor K tuned to 0,5 (see [5]).

Table 1. Accuracy, recall, precision, and average CPU times of the deterministic, robust and distributionally robust SVM models. For each formulation, the best variant in terms of accuracy is highlighted in bold.

	Deterministic (1)	Hyperrectangular Robust, (2)-(3)			Hyperellipsoidal Robust, (2)-(4)			Distributionally Robust, (5)		
$\rho_X = \rho_Y$	-	0,1	0,2	0,3	0,1	0,2	0,3	0,1	0,2	0,3
Accuracy	77,33%	**80,00**%	79,33%	74,67%	78,66%	**78,66**%	77,33%	**77,33**%	75,33%	70,67%
Recall	63,89%	**63,89**%	63,89%	58,33%	63,89%	**65,28**%	63,89%	**59,72**%	59,72%	65,28%
Precision	85,19%	**92,00**%	90,20%	84,00%	88,46%	**87,04**%	85,19%	**89,58**%	84,31%	71,21%
Avg CPU Times	2.318 sec.s	2.730 sec.s			2.768 sec.s			4.777 sec.s		

Results show that the best performing model is the robust formulation with uncertainty sets having the form of hyperrectangles (2)-(3) and $\rho_X = \rho_Y = 0{,}1$ that records an overall accuracy of 80,00%; the low value of recall (63,89%) is related to a high value of false negative rate (36,11%), while the high precision score is due to a low false positive rate (5,13%). See Tab 4. It is worth noticing

that also the ellipsoidal robust model (2)-(4) can outperform its deterministic counterpart (1) for every choice of ρ_X, ρ_Y. On the other hand the distributionally robust variant (5) shows to be the weakest one: the model records the same accuracy level of the deterministic formulation, but while its precision increases, its recall reduces. This is coherent with what concluded in [5], according to which robust classifiers are especially beneficial for low dimensionality training sets (*i.e.*, with less than approximately 500 observations), while as the training set gradually increases, features behavior is learned and higher levels of out-of-sample accuracy can be achieved via less conservative models (*i.e.*, DRO formulations). Overall, with respect to deterministic approaches, [5] proves that more conservative methods, allow finding a trade-off between increasing the average performance accuracy and protecting against uncertainty.

Considering the 15 features mentioned above, other separation models from the literature were trained using the Matlab R2022a Classification Learner App, which allows to perform automated training to seek the best classification model among more than 20 possible choices, including: *Decision Trees* (DT), linear and quadratic *Discriminant Analysis* (DA), *Logistic Regression* (LR), *Naive Bayes* (NB), *Support Vector Machines* (SVM), *k-Nearest Neighbors* (k-NN) classifiers and *Ensemble Classification* (EC). Results are reported in Table 2, where we just show accuracy levels for the sake of readability. Robust formulation with hyperrectangles (2)-(3) is able to outperform all the reported models, achieving the same accuracy level of kernel NB, and medium Gaussian SVM (both of them non-linear models).

Table 2. Accuracy levels of models from the literature trained using Matlab Classification Learner App.

	Accuracy		Accuracy		Accuracy		Accuracy
Fine DT	77,30%	Kernel NB	**80,00**%	Fine k-NN	70,70%	EC Boosted DT	76,70%
Medium DT	77,30%	Linear SVM	76,00%	Medium k-NN	72,70%	EC Subspace DA	78,70%
Coarse DT	79,30%	Quadratic SVM	76,00%	Coarse k-NN	56,70%	EC Subspace k-NN	65,30%
Linear DA	76,00%	Cubic SVM	76,00%	Cosine k-NN	76,70%	EC RUS Boosted DT	75,30%
Quadratic DA	Failed	Fine Gauss. SVM	73,30%	Cubic k-NN	72,00%		
LR	76,70%	Medium Gaussian SVM	**80,00**%	Weighted k-NN	74,70%		
Gaussian NB	Failed	Coarse Gaussian SVM	78,70%	EC Bagged DT	75,40%		

For every model, Matlab allows the internal hyperparameters tuning (*e.g.*, the maximum number of splits for a DT). Additionally, instead of manually selecting these options, the Classification Learner App also provides the hyperparameter optimization to automate the selection of these values. For a given model type, the app tries different combinations of hyperparameter values by using an optimization scheme that seeks to minimize the model classification error, and returns a model with the optimized hyperparameters. See Table 3. Overall, results show that the strongest model in term of prediction accuracy is the *Optimizable DT* (ODT), that is, a simple classifier consisting of a sequence of binary decisions hierarchically organized. The ODT has an accuracy rate of 86,00%, a recall rate of 80,56% and a precision rate of 89,23%. Its prediction rate is around 1200 observations per second and its training time is 80,125 sec-

onds. The maximum number of subdivisions considered is 6, with a subdivision criterion given by the Gini diversity index. The type of optimization considered is the Bayesian one.

Table 3. Accuracies of optimizable models trained via the Classification Learner App.

	Accuracy		Accuracy		Accuracy		Accuracy
Optimizable DA	77,30%	Optimizable NB	80,00%	Optimizable k-NN	81,30%	Optimizable EC	83,30%
Optimizable SVM	78,00%	Optimizable DT	**86,00%**				

Table 4. Confusion matrix for the hyperrectangular robust SVM formulation with $\rho_X = \rho_Y = 0{,}1$ (*left*) and confusion matrix for the ODT (*right*).

		Predicted Class	
		0	1
True Class	0	94,87%	5,13%
	1	36,11%	63,89%

		Predicted Class	
		0	1
True Class	0	91,00%	9,00%
	1	19,44%	80,56%

As opposed to black-box modeling strategies which are typically difficult to illustrate, the ODT model benefits from the great potential of interpretability and the structure we obtained is reported in Figure 1: to predict the answer (survival 0, mortality 1), we follow the decisions in the tree from the root node, down to a leaf node which contains the response. The results show that among the 15 attributes considered, the most relevant are cardiac troponin (cTnI), lactate dehydrogenase (LDH), aspartate aminotransferase (AST) and emogas (P/F). These results confirm doctors' observations. The root node considers the values of the cTnI attribute: if cTnI $\geq$ 21 and LDH $\geq$ 485,50 the patient is unlikely to survive, otherwise he has good chances of recovery. The second most important attribute to consider is LDH, while AST and P/F appear at the final level of the decision tree. Note that, for the purpose of validation, a sensitivity analysis was also carried out with respect to an increasing number of input characteristics and to different cardinalities of the cross-validation subsets (5, 10, 15, and 30) confirming the same tree structure shown in Figure 1.

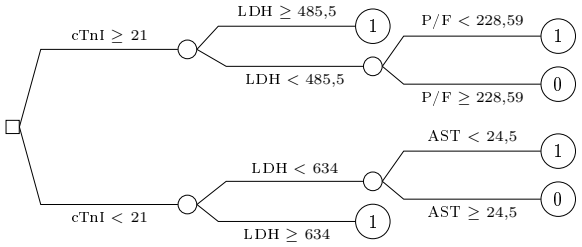

Fig. 1. The Optimizable Decision Tree.

4 Conclusions

This article presents novel data-driven optimization models to support doctors' decisions to solve one of the main problems encountered during the first months

of the health emergency: predicting and calculating the risk of mortality from COVID-19, aiming at timely identify the most appropriate assistance, diagnostic and therapeutic cares for patients. To handle the uncertainty in data features, we formulated robust and distributionally robust classificators and compared their performance with other 25 different models from the literature. Results show that the best robust classifier is the one with hyperrectangles, which outperformed all the other 25 models and recorded an 80,00% accuracy level. However, considering optimizable models available using Matlab Classification Learner App, the best results were obtained with an optimizable decision tree, recording an 86,00% accuracy and allowing the identification of the most important predictors. Future works will then consider extending our robust and distributionally robust models for such a class of optimizable decision trees. Even if this tool has not been put into practice clinically, the results confirm doctors' experience. Therefore it could be used in the future for initial patients' assessments to accurately establish the severity of his/her condition, and enabling professionals to optimize accommodation by quickly identifying patients in need of intensive care.

References

1. Aloisio, E., Chibireva, M., Serafini, L., Pasqualetti, S., Falvella, F.S., Dolci, A., Panteghini, M.: A comprehensive appraisal of laboratory biochemistry tests as major predictors of COVID-19 severity. Archives of Pathology & Laboratory Medicine **144**(12), 1457–1464 (2020)
2. Ben-Tal, A., El Ghaoui, L., Nemirovski, A.: Robust optimization, vol. 28. Princeton University Press (2009)
3. Bonetti, G., Manelli, F., Patroni, A., Bettinardi, A., Borrelli, G., Fiordalisi, G., Marino, A., Menolfi, A., Saggini, S., Volpi, R., et al.: Laboratory predictors of death from coronavirus disease 2019 (COVID-19) in the area of Valcamonica, Italy. Clinical Chemistry and Laboratory Medicine **58**(7), 1100–1105 (2020)
4. CSG of the International Committee on Taxonomy of Viruses: The species severe acute respiratory syndrome-related coronavirus: Classifying 2019-nCoV and naming it SARS-CoV-2. Nature Microbiology **5**(4), 536 (2020)
5. Faccini, D., Maggioni, F., Potra, A.F.: Robust and distributionally robust optimization models for linear Support Vector Machine. Under revision in Computers & Operations Research pp. 1–43 (2022)
6. Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., et al.: Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The Lancet **395**(10223), 497–506 (2020)
7. Lippi, G., Plebani, M.: The critical role of laboratory medicine during coronavirus disease 2019 (COVID-19) and other viral outbreaks. Clinical Chemistry and Laboratory Medicine **58**(7), 1063–1069 (2020)
8. Liu, X., Potra, F.A.: Pattern separation and prediction via linear and semidefinite programming. Studies in Informatics and Control **18**(1), 71–82 (2009)
9. Wiesemann, W., Kuhn, D., Sim, M.: Distributionally robust convex optimization. Operations Research **62**(6), 1358–1376 (2014)
10. World Health Organization: World Health Organization coronavirus disease (COVID-19) dashboard. World Health Organization (2020)