Analysis of mortality in patients with severe pneumonia

Benedetta Bertesago, Martina Fervari, Rossana Volontè

1. Introduction

Pneumonia is among the leading causes of death globally; it's the most common infective reason for admission to intensive care as well as the most common secondary infection acquired while in ICU. Among the clinical scoring systems to predict mortality in ICU, none has been specifically developed for patients with severe pneumonia. However, an early and accurate prediction of patients' mortality is crucial for optimizing treatment plans and potentially improving survival rates. The goal of this project is to develop a model to predict the mortality of ICU patients with severe pneumonia within the next 24 hours, to assist healthcare providers in making critical decisions about intervention strategies.

2. Material and Methods

The CarpeDiem dataset features 585 critically ill patients, corresponding to 12495 patient-ICU-days, who required mechanical ventilation and underwent a bronchoalveolar lavage (BAL) exam given suspicion for severe lung infection. Patients were categorized depending on the type of pneumonia by which they were affected (COVID-19, other respiratory viruses, bacterial, respiratory failure unrelated to pneumonia) and 45% of them had an unfavorable outcome (discharge to hospice or death). For each patient, the dataset presents information about demographics (age, ethnicity, gender, race, BMI, smoking status, admission source, admission SOFA and APS scores) and outcomes (discharge disposition, cumulative ICU days, number of ICU stays, cumulative intubation days, tracheostomy requirement). Moreover, it contains daily clinical parameters values, including vital signs (temperature, heart rate, systolic and diastolic blood pressure, respiratory rate, oxygen saturation, urine output), laboratory parameters (white blood cells, neutrophils, lymphocytes, hemoglobin, platelets, bicarbonate, creatinine, albumin, bilirubin) and mechanical support devices (intubation, hemodialysis, CRRT, PEEP). It also includes information about BALs and clinical pneumonia episode adjudication data: category (CAP, HAP, VAP, non-pneumonia control), etiology (bacterial, viral, culture-negative), duration.

Regarding the preprocessing phase, we manipulate the dataset to obtain a uniform sample of patients, following these steps:

- We remove patients with at least one *ICU_stay* longer than 31 days, people who are under 30 or over 80 years old or who have a *BMI* greater than 60.
- We remove the features presenting a percentage of null values higher than 50%, assuming they are missing at random; in particular, the removed variables are: *Global_cause_failure, CRP, D_dimer, Ferritin, LDH, Lactic_acid* and *Procalcitonin*.
- We analyze features related to mechanical ventilation or drug administration, which present some values missing not at random when the patient is not intubated or doesn't receive medicine. For this reason, we substitute null values with 0 in *Norepinephrine_rate* when *Norepinephrine_flag* is 0 and in *PEEP, FiO2, Plateau_Pressure, Lung_Compliance, PEEP_changes, FiO2_changes* and *Respiratory_rate_changes* when *Intubation_flag* is 0.
- Regarding the BAL exam, we decide to remove *Episode_is_cured* due to the lack of interpretability of this variable, whereas, to take into account the results of the BAL, for each patient we replicate the other available information in the days following the exam.
- We remove the row associated to the last day of patients who died, since these values could not be used by the model for the prediction. Moreover, we check the percentage of null values for each patient and each ICU stay, in particular looking at values registered in the first and

last day of the stay. Since this percentage is always smaller than 30%, we keep the observations unmodified.

- We proceed with Feature Engineering, creating some variables to keep track of the clinical history of patients. Specifically, we introduce *Cumulative_days*, *Cumulative_intub_days* and *Cumulative_bal*, which progressively count the number of days spent in ICU, the number of days spent intubated and the number of BAL exams performed for each patient. We create *Bacterial_duration* feature representing the decreasing duration of a bacterial pneumonia episode, discovered through a BAL exam. We design also the variable *GCS_total*, as the sum of values of *GCS_eye_opening*, *GCS_motor_response* and *GCS_verbal_response*. Finally, we construct the target variable *Next_day_outcome*, setting it to 1 if the patient dies in the next 24 hours, 0 otherwise.
- We study the correlation among numerical features to prevent problems of collinearity inside the models. In particular, we consider the pairs of variables that present a correlation coefficient higher than 0.72 in absolute value.
- We check the range of variables, to eventually discard observations incorrectly registered and so clinically meaningless.

Afterwards, we encode the categorical features using a OneHotEncoder to represent them as dummy variables and we split the final dataset into a training and a test sets, stratifying the split with respect to the variable *Next_day_outcome* and keeping a proportion of 20% of observations to evaluate the final model. Before the fitting of the model, we substitute the remaining null values using a KNNImputer, choosing the best number of neighbours through Cross Validation by minimizing the MSE between some masked values and their corresponding imputations. After scaling the data, we proceed with the construction of a suitable classification model for the prediction of the target variable. In particular, we exploit LogisticRegression, SGDClassifier, SVC and RandomForest, fine-tuning the parameters to maximize the balanced accuracy through 5-Fold Cross Validation. Moreover, due to the imbalance between the 2 classes, we experiment the SMOTE technique on RandomForest to enrich the training dataset with new observations of the minority class obtained as combinations of existing samples. We compare the obtained models by assessing their performances and choose the best one among them, on which we perform Feature Selection. Specifically, we deploy SelectFromModel function to choose the most significant variables for the model. Then, we follow two alternative approaches to further refine the predictive model: SequentialFeatureSelector function, which is a greedy-search algorithm that carries out a backward selection of the features, and SelectKBest function, which evaluates the significance of each variable individually with respect to ANOVA F-statistic and Mutual Information Coefficient. Finally, the reduced models are evaluated on the test set to understand their ability in predicting especially the minority class, namely if the patient will die in the following 24 hours.

3. Results

After the preprocessing analysis on *ICU_stay*, *Age* and *BMI* distributions and the removal of variables *Global_cause_failure*, *CRP*, *D_dimer*, *Ferritin*, *LDH*, *Lactic_acid*, *Procalcitonin* and *Episode_is_cured*, we obtain a reduced dataset, consisting in 426 patients, for a total of 6209 rows and 65 columns. Considering also the variables introduced with Feature Engineering, we study the correlation among 44 numerical features. Specifically, we examine the pairs of variables that present a correlation coefficient higher than 0.72 in absolute value and for each couple we keep only the feature which has fewer missing values. In particular:

- *WBC_count* and *Neutrophils* are highly correlated (with correlation coefficient 0.904) and so only the first one is retained.
- *RASS_score*, *GCS_total*, *GCS_eye_opening*, *GCS_motor_response* and *GCS_verbal_response* are characterized by the following correlation matrix:

	$RASS_{score}$	GCS_{total}	$GCS_{eye\_opening}$	$GCS_{motor\_response}$	$GCS_{verbal\_response}$
$RASS_{score}$	1.00	0.781	0.773	0.773	0.486
GCS_{total}		1.00	0.872	0.911	0.779
$GCS_{eye\_opening}$			1.00	0.761	0.530
$GCS_{motor\_response}$				1.00	0.507
$GCS_{verbal\_response}$					1.00

We decide to keep only *GCS_total*, since it has the highest correlation with all the other four variables.

- *Bicarbonate* and *ABG_PaCO2* present a correlation coefficient of 0.747 and so only *Bicarbonate* is retained.
- *Diastolic_blood_pressure* and *Mean_arterial_pressure* present a correlation coefficient of 0.731 and so only *Diastolic_blood_pressure* is retained.
- *PEEP*, *FiO2* and *Plateau_Pressure* are highly correlated and characterized by the following correlation matrix:

	PEEP	*FiO2*	*Plateau_Pressure*
PEEP	1.00	0.738	0.840
FiO2		1.00	0.825
Plateau_Pressure			1.00

We keep only *FiO2* as feature of the predictive model, since it presents fewer missing values.

- *Cumulative_days* and *Cumulative_intub_days* present a correlation coefficient of 0.915 and so only *Cumulative_intub_days* is retained since it's assumed to be more relevant for the analysis on pneumonia.

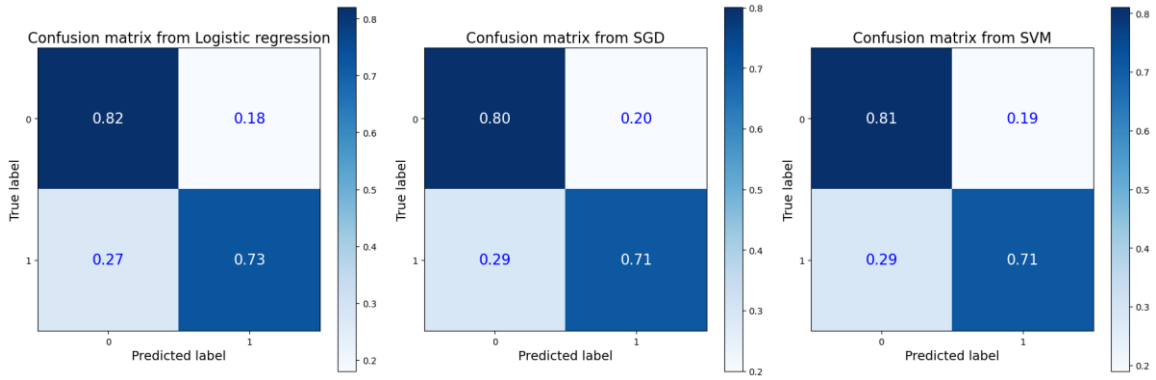
After the encoding of categorical variables, we notice that the minority class represents just 3% of total observations, so the splitting of the dataset is performed maintaining this proportion in both train and test sets. Therefore, in order to take into account this strong imbalance, we proceed in our analysis introducing some metrics and techniques that weight the predictions based on the frequencies of the two classes. Regarding the imputation of null values, by means of a GridSearch algorithm we select $k = 14$ as best number of neighbours for the KNNImputer. Then, considering the features retained so far, we implement the following classification models, predicting as output the possibility that a patient will die the following day, represented by the variable *Next_day_outcome:*

- Logistic Regression, fixing the penalty to 'L1' and the regularization coefficient C to 0.1.
- SGDClassifier, fixing the loss function to 'hinge' and the regularization coefficient α to 0.65.
- SVC, fixing the regularization coefficient C to 0.2 and considering a linear kernel.

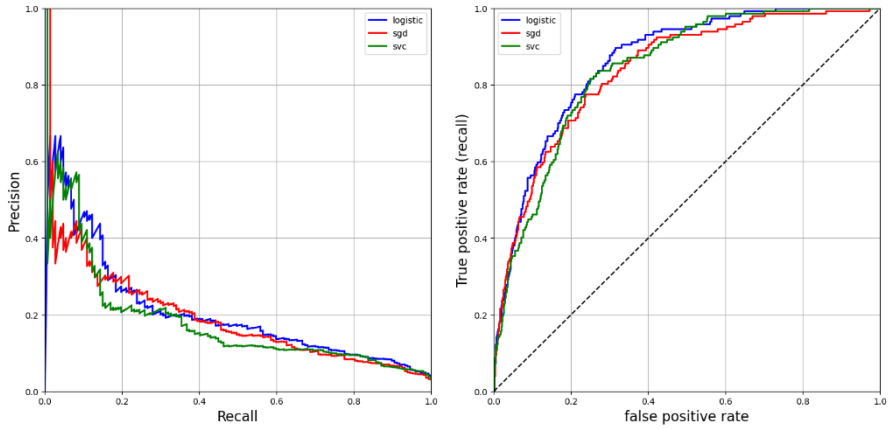
The parameters are found through a Grid Search algorithm, maximizing the balanced accuracy score, defined as the average between sensitivity and specificity. By means of 5-Fold Cross Validation, we compute the confusion matrix and some corresponding scores to evaluate each model individually on the training set.

	LogisticRegression	**SGDClassifier**	**SVC**
Accuracy	0.817	0.798	0.807
Balanced Accuracy	0.771	0.751	0.761
F1 score	0.196	0.177	0.185
Precision	0.114	0.101	0.107
Recall	0.723	0.702	0.712

Table 1: Performance of the models with 5-Fold CV

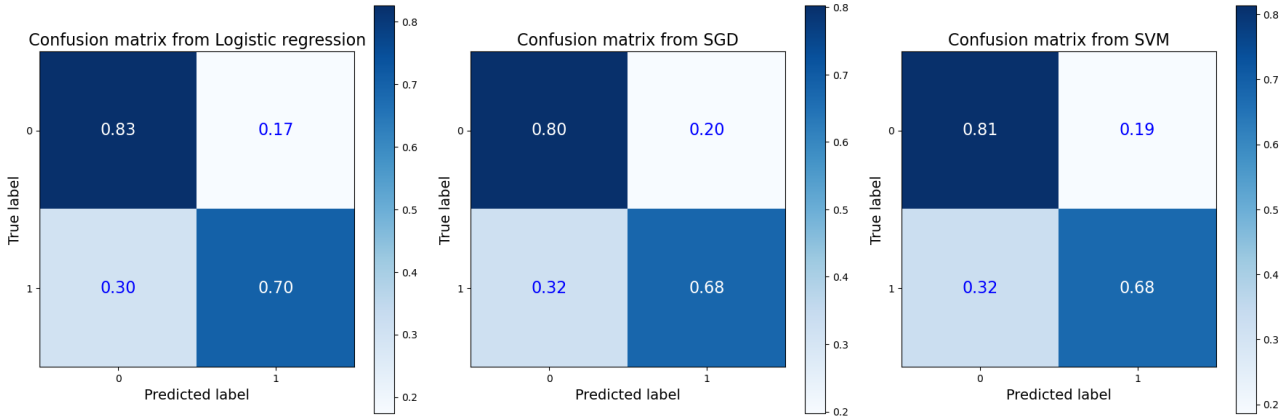


Moreover, in order to compare these classifiers, we look at the Precision-Recall and ROC curves and we evaluate the models on the test set, selecting the Logistic Regression model as the most powerful in predicting both classes correctly.



	AUC
LogisticRegression	0.864
SGDClassifier	0.841
SVC	0.844

Table 2: AUC with 5-Fold CV scores



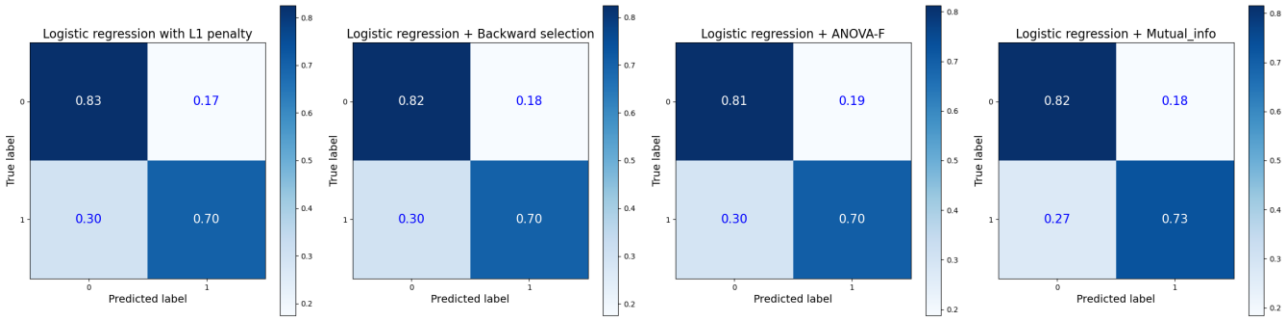
Additionally, we try to implement another classification model through RandomForest, obtaining only unsatisfying results, regardless of the choice of the parameters. Therefore, supposing the strong imbalance as the main cause of the poor performance, we exploit the SMOTE technique, trying to increase the number of observations of the minority class. Proceeding in this way, we apply a Grid Search algorithm to find the best configuration for the RandomForest parameters and test the final model on the test set, still obtaining poor results.

The last step of the analysis is the Feature Selection performed on the Logistic Regression classifier. In particular, since the model is fitted using L1 penalty, the function SelectFromModel automatically removes the features characterized by a null coefficient, keeping in the end just 50 variables out of 79. Successively, we apply the function SelectKBest and backward selection on the remaining features, identifying the number of variables maximizing the balanced accuracy. Specifically, the first

method selects 30 best features when using ANOVA F-statistic and 37 when using Mutual Information coefficient, whereas the second approach retains 49 variables. In order to find the best reduced model, we evaluate the mentioned classifiers on the test set, computing the following metrics:

	Logistic	BS	ANOVA-F	Mutual Info
Accuracy	0.822	0.821	0.809	0.813
Balanced Accuracy	0.764	0.764	0.758	0.773
F1 score	0.196	0.195	0.186	0.195
Precision	0.114	0.114	0.107	0.113
Recall	0.703	0.703	0.703	0.730
AUC	0.764	0.764	0.758	0.773

Table 3: Performance of the reduced models on the test set, considering Backward Selection(BS) and SelectKBest function



4. Discussion and conclusion

By looking at the results of the implemented classifiers on the test set, we can observe that Logistic Regression Model with Mutual Information coefficient outperforms the other reduced versions and experimented models in the prediction of the classes. Hence, we deduce that the feature selection improves the performance of the model, making it convenient since it is less data demanding. In particular, the final model identifies these significant features:

General Info & BAL		Vital signs		Laboratory Parameters		Mechanical Ventilation	
ICU_{day}	0.438	$SOFA_{score}$	0.251	ABG_{pH}	-0.179	$Norepinephrine_{rate}$	0.197
$COVID_{True}$	-1.647	$Hemodialysis_{flag}$	-0.597	ABG_{PaO2}	-0.350	$Norepinephrine_{flag}$	-0.356
$Male$	0.179	$CRRT_{flag}$	-0.405	WBC_{count}	0.227	$Respiratory_{rate}$	0.106
$OtherPneumonia$	0.238	$Temperature$	-0.137	$Lymphocytes$	0.012	$FiO2$	0.319
$NeverSmoker$	0.044	$Heart_{rate}$	0.276	$Hemoglobin$	-0.042	$LungCompliance$	0.043
$Episode_{duration}$	-0.130	$Systolic$	-0.066	$Platelets$	-0.496	$PEEP_{changes}$	-0.175
$Cumulative_{BAL}$	0.173	$Diastolic$	-0.046	$Bicarbonate$	-0.047	$Respiratory_{rate\_changes}$	-0.169
$Bacterial_{duration}$	-1.542	$UrineOutput$	-0.169	$Albumin$	-0.129		
$Etiology_{Bacterial/Viral}$	0.859	$OxygenSaturation$	-0.177	$Bilirubin$	0.021		
$Etiology_{Before-first-BAL}$	-3.190	GCS_{total}	-0.470				
$Etiology_{Viral}$	-0.344						

Table 4: Coefficients of Logistic Regression Classifier with Mutual Information

Coefficients with positive sign correspond to variables for which a higher value is associated with a larger probability of patient's death, while coefficients with negative sign indicate that higher values of these features imply a lower risk of mortality.

Overall, the classifier is quite effective and efficient in predicting patients' mortality and shows especially the importance of performing BAL exams and blood tests to monitor the clinical condition of patients.