UDACITY

Machine Learning Engineer Nanodegree

Capstone Project

Predicting the need for mechanical ventilation using the MIMIC-III database

Leonardo Watanabe Kume

October 30, 2020

1 Definition

1.1 Project Overview

The start of this decade will be forever marked by the COVID-19 (Coronavirus Disease 2019) pandemic, it has amassed more than one million deaths worldwide until October of 2020 and affected every person somehow. Caused by the SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2) virus, it enters the body through the airways and infects primarily the respiratory system, causing flu-like symptoms, such as: cough, fever and difficulty breathing. Most people present a mild case, but around 20%[1] of the infected ones require hospitalization. Due to high infectivity, COVID-19 spread quickly and infected hundreds of thousands in a matter of weeks, causing health services to be overcrowded. In this scenario, ventilators emerged as one of the key resources in the management of the pandemic and its shortage led doctors having to choose between patients who would be saved and who wouldn't.

The goal of this project is to use MIMIC-III[2] data to develop a ML (Machine Learning) model capable of predicting whether an admitted patient will require mechanical ventilation. It will not only give doctors an early estimate of the demand for ICU beds and ventilators, but also allows them to save more lives by distributing the resources efficiently.

1.2 Problem Statement

With the rise of machine learning, health services are beginning to adopt ML-powered EWS (Early Warning Systems) to anticipate sepsis, clinical deterioration, readmission and mortality. Bearing in mind that one of the problems that lead to unnecessary deaths is the delay in treatment, the COVID-19 pandemic presents a great opportunity to develop ML models that can warn the medical team of the potential need for mechanical ventilation with hours of advance.

1.3 Metrics

In medicine, most of the problems involve imbalanced datasets and this one is no exception, since the rate of intubation for COVID-19 patients is approximately 3%[3]. In this context, judging an ML model by accuracy and AUC-ROC (Area Under the Receiver Operating Characteristic Curve) can be misleading, since both metrics can show satisfactory results by simply predicting every example as the majority class. Better metrics for these situations are PPV (Positive Predictive Value), NPV (Negative Predictive Value), sensitivity, and specificity.

PPV is defined by the number of true positives divided by the sum of true positives and false positives (Equation 1). In a practical situation, it would tell us the probability of an alerted patient needing MV in the near future.

$$PPV = \frac{True\ positives}{True\ positives + False\ positives} \quad (1)$$

NPV is defined by the number of true negatives divided by the sum of true negatives and false negatives (Equation 2). In a practical situation, it would tell us the probability of an unalerted patient not needing MV in the near future.

$$NPV = \frac{True\ negatives}{True\ negatives + False\ negatives} \quad (2)$$

Sensitivity is defined by the number of true positives divided by the total number of examples from the positive class (Equation 3). A low sensitivity tells us that the model has a high rate of false negatives, not alerting patients that would need MV in the near future.

$$Sensitivity = \frac{True\ positives}{True\ positives + False\ negatives} \quad (3)$$

Specificity is defined by the number of true negatives divided by the total number of examples from the negative class (Equation 4). A low specificity tells us that the model has a high rate of false positives, alerting too many patients unnecessarily.

$$Specificity = \frac{True\ negatives}{True\ negatives + False\ positives} \quad (4)$$

2 Analysis

2.1 Data Exploration

The MIMIC-III (Medical Information Mart for Intensive Care) dataset contains data from 46,520 adult patients admitted to critical care units of the Beth Israel Deaconess Medical Center in Boston, Massachusetts[2]. There's a wide variety of data available in the MIMIC-III database but, for this project, the focus was on commonly available measurements, such as: age, sex, heart rate, temperature, mean arterial pressure (MAP), blood urea nitrogen, hematocrit, etc.

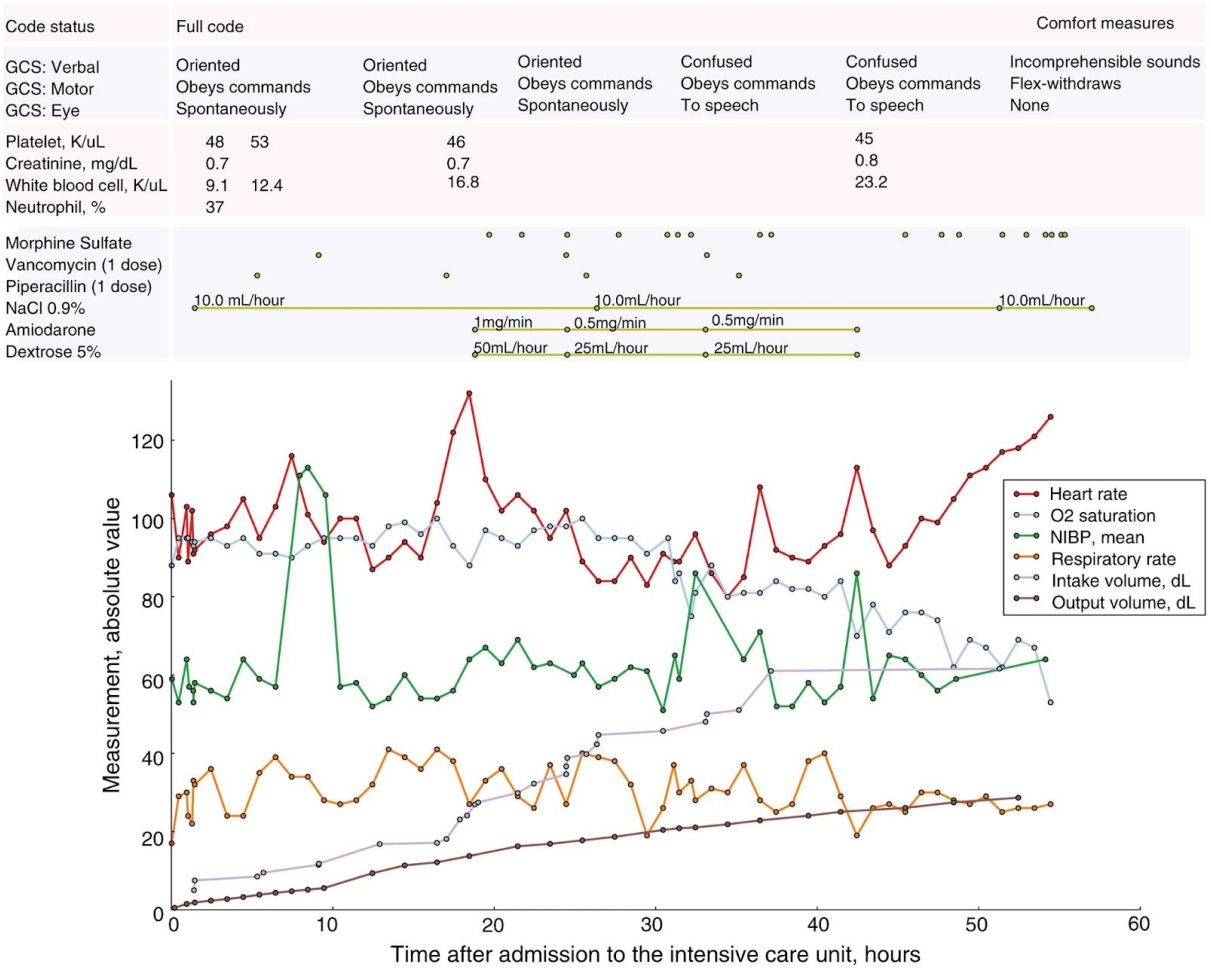

Figure 1 - Example of extracted variables for a single patient from the MIMIC-III database. (Source: Johnson, 2016).

2.2 Exploratory Visualization

An example of the vital signs and a few laboratory results from an MV and a non-MV patient are shown in Figure 2.

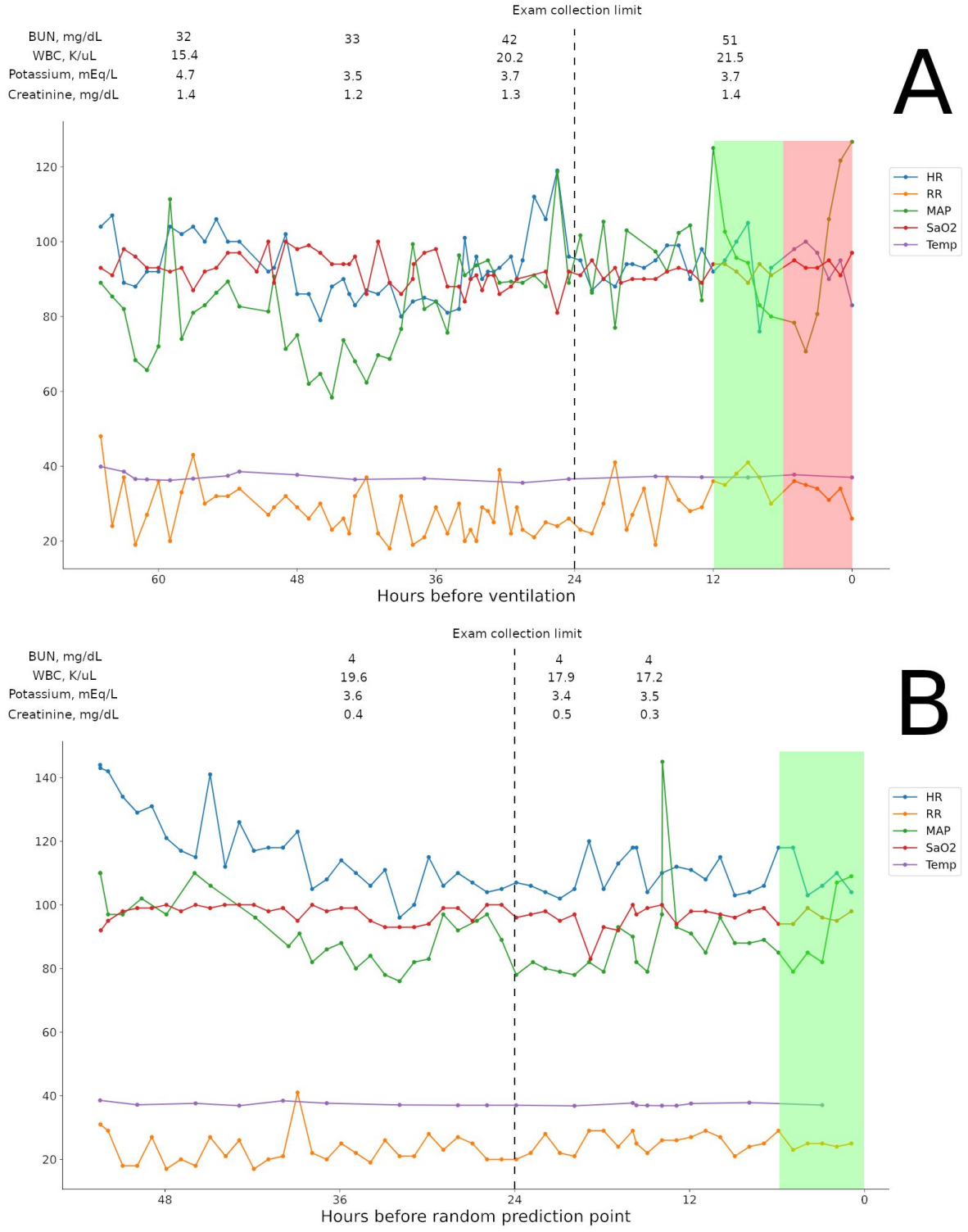

Figure 2 - Example of data collected from an MV (A) and a non-MV (B) patient. Highlighted in red is the discard zone, and in green, the collection window. The dashed line represents the limit of time where a laboratory result can be considered for the analysis.

From Figure 2 is possible to observe some differences between these two patients. The MV patient presents much more unstable vital signs and worsening laboratory results.

2.3 Algorithms and Techniques

Since this is a binary classification problem from a tabular dataset, there's a wide variety of algorithms that might perform well. Taking that into consideration, various algorithms were tested to check which one is the best for this particular problem. The algorithms tested were: LR (Logistic Regression), ComplementNB (Complement Naïve Bayes), DTC (Decision Tree Classifier), SVM (Support Vector Machine) with a linear kernel, and XGBoost (Extreme Gradient Boosting Machine).

2.4 Benchmark

When it comes to predicting mechanical ventilation, the academic literature is scarce. CURB-65[4] (Confusion, Urea, Respiratory Rate and age above 65) and PSI[5] (Pneumonia Severity Index) are two clinical prediction rules widely used to predict 30-day mortality in patients with CAP (Community-Acquired Pneumonia) and they can also be used to predict MV; The issue is that they are applied only during admission, which means that it is not applicable in the context of this project.

The APACHE-II[6] score is a commonly used tool to assess severity and predict mortality in ICU patients and it takes into consideration a wide variety of variables, from the first 24h of admission to the ICU, and assigns points according to their amendments- higher scores correspond with increased severity and chance of death. Even though it is not designed for the specific context of this project, it is interesting to check the performance of existing tools. For this project, acute renal failure, alveolar-arterial gradient and history of severe organ insufficiency or immunocompromised were not available to calculate the final score.

Since there is no other real-life benchmark for this experiment, the chosen algorithm was benchmarked against the default implementation of the other algorithms described in the previous section.

3 Methodology

3.1 Data Preprocessing

AWS Athena was used to query the MIMIC-III tables and assemble the views necessary for this project. Then the subsequent analysis was done using Python 3.8.5.

3.1.1 Mechanical Ventilation Data

In order to label the patients that were mechanically ventilated, a SQL script was available on the MIMIC-III GitHub page[7]. It was validated in a peer-reviewed paper[8].

3.1.2 Cohort Definition

The initial cohort was defined as patients who are older than 18 years, that did not have secondary stays and had a length of stay longer than 24 hours. After applying these exclusion criteria, 19.331 patients were selected from the 46.520 original ones. Vital signs, laboratory results, ICD-9 and mechanical ventilation data were subsequently extracted from those patients.

Patients that were mechanically ventilated in less than 24 hours after admission to the ICU were excluded because they wouldn't have enough data for the analysis. Moving along, patients that did not have at least 5 measurements of each vital sign in the prediction period were also excluded.

Finally, only patients that had an ICD-9 code indicating the presence of respiratory disease (ICD-9 460-519) were selected for the final cohort. Leaving 3,358 patients, of whom 841 (33.4%) were mechanically ventilated.

3.1.3 Prediction Point and Collection Window Definition

Since the outcome of interest is MV, the point of prediction was defined as 6 hours before the event and the collection window is 6 hours (highlighted in green in Figure 2) before the prediction point. That means that all vital signs and lab results after that point were excluded (highlighted in red in Figure 2).

3.1.3 ICD-9

The ICD-9 codes were used to define binary features that indicate whether a particular patient presents diseases from the following categories, which were defined according to the comorbidities used by the PSI:

- Neoplasms
- Heart Failure
- Cerebrovascular diseases
- Hepatic disease
- Renal disease

3.1.4 Vital Signs and Laboratory Tests

The present outliers in vital signs and labs were treated using IQR (Interquartile Range) (Equation 5). Lower and upper limits were defined by multiplying the IQR by a factor of 5 (Equations 6 and 7).

$$IQR = Q3 - Q1 \quad (5)$$

$$Upper\ limit = Q3 + IQR * 5 \quad (6)$$

$$Lower\ limit\ =\ Q1 - IQR * 5 \quad (7)$$

Where:

- Q1 = First quartile
- Q3 = Third quartile
- IQR = Interquartile Range

For variables that the calculated lower limit was negative, it was manually set as zero, and for Oxygen Saturation the upper limit was set as 100.

3.1.4.1 Features calculation

Vital signs were individually collected from the prediction window (6h). Then the minimum and maximum values from that time frame were selected as features.

Since laboratory tests are not as frequent as vital signs collections, the collection period was extended to 24 hours, and the most recent value was subsequently selected as a feature.

Lastly, there were 23 features available to train the model, which are described in Table 1.

Vital Signs	Laboratory Results
• Max. Heart Rate • Min.Heart Rate • Max.Respiratory Rate • Min. Respiratory Rate • Max.Temperature • Min. Temperature • Min. MAP (Mean Arterial Pressure) • Max. MAP (Mean Arterial Pressure) • Min. Oxygen Saturation • Min. GCS (Glasgow Coma Scale)	• Sodium • Potassium • White Blood Cell Count • Creatinine • Hematocrit • Glucose • Blood Urea Nitrogen • Bicarbonate
ICD-9	**Demographics**
• Neoplasms • Heart Failure • Cerebrovascular diseases • Hepatic disease • Renal disease	• Gender • Age

Table 1 - Features available to train the model.

3.2 Implementation

Sklearn (version 0.23.2) was used for the implementation of this project.

3.2.1 Defining the initial dataset

The initial dataset was defined using all of the 23 features and then excluding the rows that had any missing values. This resulted in 2,298 patients, where 597 (30%) were of the positive class. Normalization was performed using the MinMaxScaler algorithm, normalizing all the features to a range of [-1, 1].

3.2.2 Testing the algorithms

Initially, the algorithms mentioned in the Algorithms and Techniques section were validated using Repeated K-Fold Cross Validation, with 10 folds and 3 repetitions, using the default hyperparameters and balanced class weights. The results are presented in Table 2.

Algorithm	ROC-AUC	F1	Recall	Precision
Naive Bayes	0.673±0.036	0.435±0.051	0.406±0.059	0.476±0.068
DTC	0.551±0.037	0.337±0.057	0.340±0.064	0.336±0.058
Random Forest	0.694±0.035	0.180±0.067	0.107±0.044	0.609±0.147
Logistic Regression	0.689±0.034	0.480±0.042	0.599±0.059	0.403±0.042
SVM	0.688±0.034	0.481±0.041	0.595±0.059	0.406±0.042
XGBoost	0.679±0.031	0.389±0.040	0.338±0.050	0.464±0.044

Table 2 - Results of the performance from various algorithms using the default hyperparameters and balanced class weights.

With quite identical performance, both the Logistic Regression and SVM algorithms were chosen to be further refined.

3.3 Refinement

This section describes the refinement process of the SVM and LR models. The dataset was split into training, validation and test sets, using a proportion of 60/20/20 respectively.

3.3.1 Hyperparameter tuning

Hyperparameter tuning was performed on the validation set by Grid Search, using the default 5-fold cross-validation and recall as the scoring metric. The search grids for the SVM and LR models are described below.

```
params_grid = {
    'loss': ['hinge', 'squared_hinge'],
    'dual': [True, False],
    'tol': [0.01, 0.1, 1],
    'C': [0.1, 1, 10],
    'class_weight': [{0:1, 1:1}, {0:1, 1:2}, {0:1, 1:3}]
}
```

```
params_grid = {
    'dual': [True, False],
    'tol': [0.01, 0.1, 1],
    'C': [0.1, 1, 10],
    'fit_intercept':[True, False],
    'solver':['lbfgs', 'liblinear'],
    'class_weight': [{0:1, 1:1}, {0:1, 1:2}, {0:1, 1:3}]
}
```

The range of values was defined manually to avoid the overclassification of one of the classes. The goal was to give the model the best opportunity to achieve a balance between sensitivity and specificity metrics.

3.3.2 Feature selection

First, the models were trained using the 25 features and a training, validation and test split of 60/20/20 respectively. The results are presented below in Table 3.

		Metric				
Model	Dataset	Sensitivity	Specificity	PPV	NPV	F1-score
SVM	Train	64.43%	68.25%	41.74%	84.46%	0.51
	Test	59.54%	68.10%	42.86%	80.73%	0.50
LR	Train	62.46%	67.48%	38.94%	84.41%	66.23%
	Test	52.67%	68.40%	40.12%	78.25%	63.89%

Table 3 - Results from training the models with the full 23 features.

Hyperparameters:

```
LinearSVC(C=10, class_weight={0: 1, 1: 3}, loss='hinge',
random_state=42, tol=0.1)
```

```
LogisticRegression(C=0.1, class_weight={0: 1, 1: 3},
random_state=42, tol=0.01)
```

From the feature importance plots (Figure 3), it is possible to observe that there are some features that don't contribute to the classification, only serving as noise. Another thing to notice are features whose contribution doesn't make sense. For example, the minimum heart rate, contributes to the positive classification, while the expected behavior would be the opposite. The same can be said about the minimum MAP and minimum temperature. This might indicate that these particular features are being overfitted to the dataset.

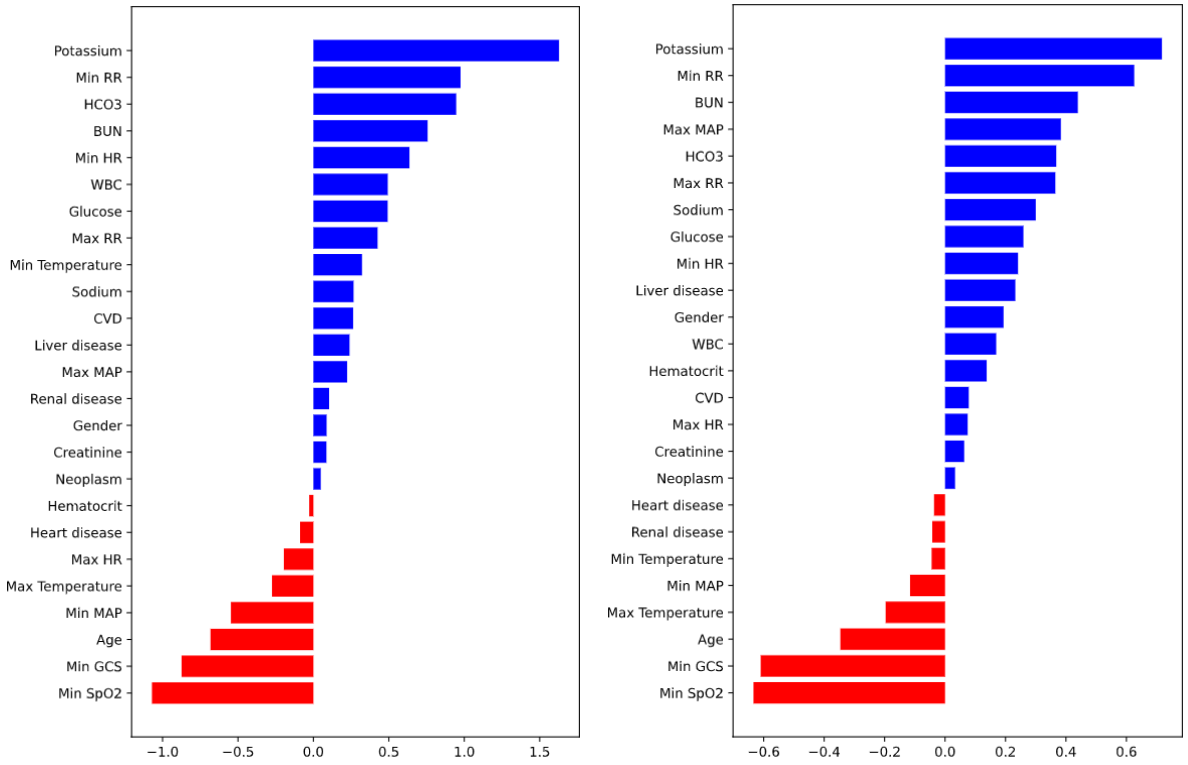

Figure 3 - Feature importance from the SVM and LR models.

After taking these points into consideration, the following features were selected:

1. Maximum Respiratory Rate
2. Bicarbonate
3. White Blood Cell Count
4. Potassium
5. Blood Urea Nitrogen
6. Maximum Heart Rate
7. Maximum Mean Arterial Pressure
8. Presence of Liver disease
9. Hematocrit
10. Glucose

11. Creatinine
12. Sodium
13. Minimum Oxygen Saturation
14. Minimum GCS

4 Results

4.1 Model Evaluation and Validation

After removing 11 features, the models showed an improvement in performance, which that the removed ones were indeed only adding noise to the classification.

		Metric				
Model	Dataset	Sensitivity	Specificity	PPV	NPV	F1-score
SVM	Train	60.58%	69.60%	40.19%	83.96%	0.48
	Test	58.78%	69.94%	44.00%	80.85%	0.61
LR	Train	63.35%	66.54	39.61%	83.98%	0.49
	Test	62.60%	66.87%	43.16%	81.65%	0.63

Table 4 - Results for the final models trained with the selected 14 features.

Hyperparameters:

```
LinearSVC(C=1, class_weight={0: 1, 1: 3}, loss='hinge',
random_state=42, tol=1)


LogisticRegression(C=0.1, class_weight={0: 1, 1: 3},
random_state=42, tol=1)
```

Three repetitions of 10-fold cross-validation were performed on the final models to calculate the 95% CI (Confidence Interval) for the sensitivity, PPV, AUC-ROC and F1-score (Table 4).

Model	Sensitivity	PPV	AUC-ROC	F1-Score
SVM	0.554-0.676	0.352-0.434	0.610-0.724	0.440-0.514
LR	0.574-0.696	0.351-0.429	0.655-0.725	0.447-0.527

Table 5 - 95% CI from repeated K-fold cross-validation results.

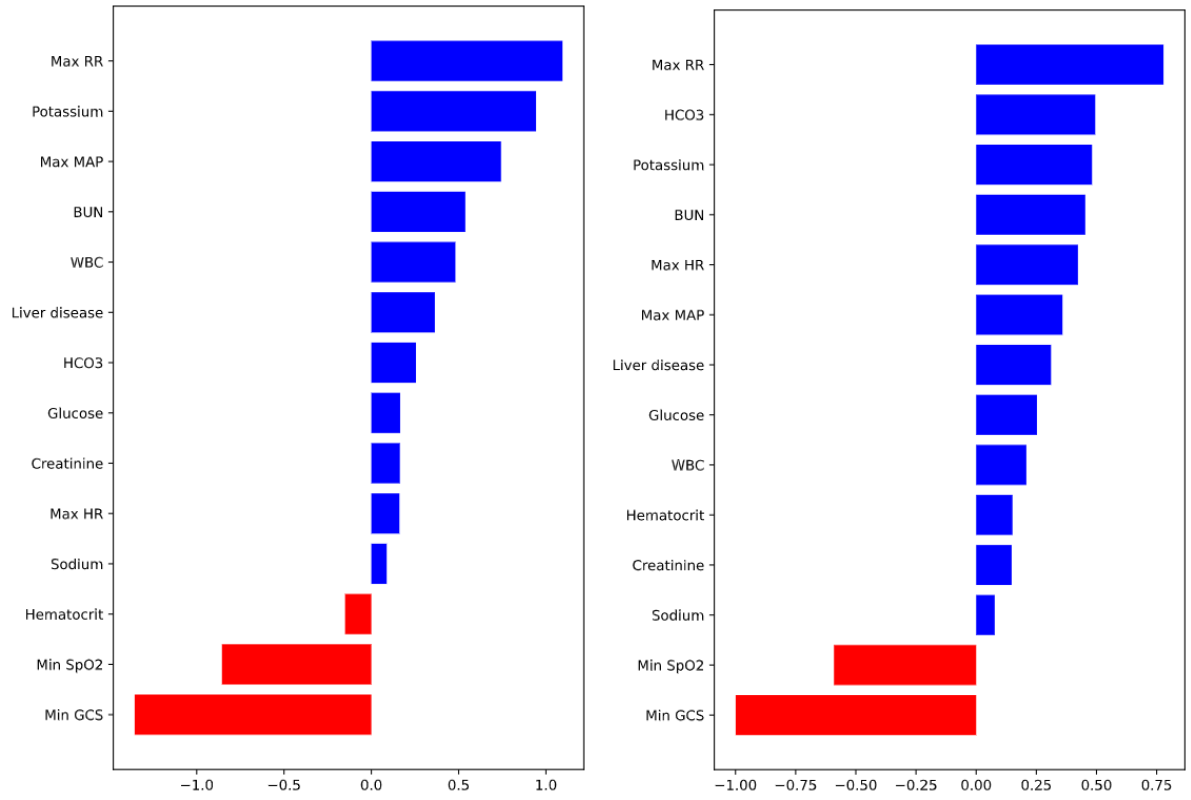

Figure 4 - Feature importance from the 14 selected features.

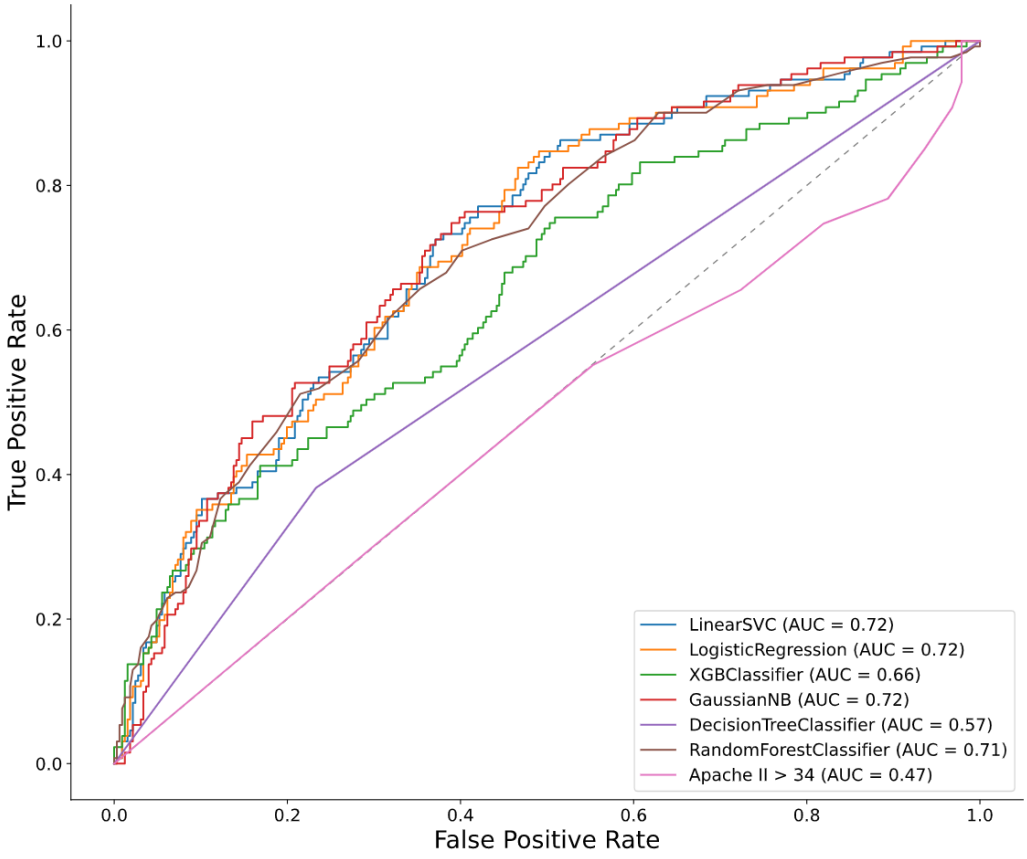

Figure 5 - Plot showing a comparison from the AUC-ROC of the base, the fine-tuned models and an Apache II score higher than 34.

From the AUC-ROC alone, it seems that the fine-tuned models showed no improvement when compared to the base RF and NB models, but by taking a closer look at the metrics we can see a clear difference.

Model	Sensitivity	Specificity	PPV	NPV
SVM	58.78%	69.94%	44.00%	80.85%
LR	62.60%	66.87%	43.16%	81.65%
RF	10.69%	76.69%	39.68%	75.53%
NB	37.40%	87.12%	53.58%	77.60%

Table 6 - Metric comparison from models with similar AUC-ROC.

4.2 Justification

Comparing the results from the ML models to the Apache II score, is not really fair, since it was not designed to predict the need for mechanical ventilation in the near future, but the stark difference in the performances shows that ML models have a lot of potential.

One of the challenges when developing ML models for medicine is finding the right balance between sensitivity and specificity on imbalanced datasets. Since classifying a sick patient as not sick (False Negative, type II error) is much more severe than the opposite (False Positive, type I error), these types of errors are penalized when training a model. But, since there isn't a strict definition of what the right balance is, it's up to the data scientists to decide.

In this project, two algorithms with similar base performances were fine-tuned to detect approximately 60% of patients admitted to the ICU that would require mechanical ventilation in the next 6 hours. By achieving an AUC-ROC of 0.72 (95% CI 0.655-0.725), a PPV of 43,16% (95% CI 0.351-0.429), a sensitivity of 62.60% (95% CI 0.574-0.696) and an F1-score of 0.63 (95% CI 0.447-0.527), the Logistic Regression algorithm outperformed SVM by a small margin.

Using only easily available features, this project shows the potential that ML models have to predict the need for MV in the next 6 hours. After the COVID-19 pandemic showed the world that shortage of ICU beds and ventilators cause unnecessary deaths, the usage of ML in the next pandemic might assist in saving thousands of lives.

References

1. World Health Organization. Knowing the risk for COVID-19 (2020). https://www.who.int/indonesia/news/detail/08-03-2020-knowing-the-risk-for-covid-19

2. Johnson, A. *et al.* MIMIC-III, a freely accessible critical care database. *Sci Data* **3,** 160035 (2016). https://doi.org/10.1038/sdata.2016.35

3. Meng, L. *et al.* Intubation and Ventilation amid the COVID-19 OutbreakWuhan's Experience. *Anesthesiology* **132**, 1317–1332 (2020). https://doi.org/10.1097/ALN.0000000000003296

4. Fine, M.J. *et al.* A Prediction Rule to Identify Low-Risk Patients with Community-Acquired Pneumonia. *N Engl J Med* **336**, 243–250 (1997). https://doi.org/10.1056/NEJM199701233360402

5. Lim, W.S. *et al.* Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study. *Thorax* **58**, 377–382 (2003). https://doi.org/10.1136/thorax.58.5.377

6. Knaus, W. A., Draper, E. A., Wagner, D. P. & Zimmerman, J. E. APACHE II: A severity of disease classification system. *Critical Care Medicine* **13**, 818–829 (1985). https://doi.org/10.1097%2F00003246-198510000-00009

7. Johnson, A. SQL scripts to define patients that require mechanical ventilation from the MIMIC-III database. Available at: MIMIC-III GitHub

8. Hsu, D. J. *et al.* The Association Between Indwelling Arterial Catheters and Mortality in Hemodynamically Stable Patients With Respiratory Failure. *Chest* **148**, 1470–1476 (2015). https://doi.org10.1378/chest.15-0516