SPH6004 AY22/23 Semester 2. Assignment 1

By: Joseph Liew

Report Title: Development of a Linear, Non-Linear and Ensemble Regression Models to predict intubation outcomes in an ICU

(Note: Tables and Figures are located at the report end.)

Aims

1) Develop and implement predictive models for intubation outcomes in an ICU on a static structure dataset.
2) Develop and implement feature selection methods to enhance prediction models.
3) Analyse and explain the differences in performance from different models.

Strategy

The ICU is a high-stake environment challenged by tight time and resources. Hence, any useful prediction model for intubation outcome should use readily available measurements that have high prediction performance.

1.1) Strategy Outline

Broadly, the strategy is to start with initial models to testing possible model patterns or features. Then, using information gained, the model designs were gradually tweaked and features selected. This was done through repeated cycles of hyperparameter tuning, feature selection, model training, testing and validation. Progressively, weaker models were weeded out, retaining the model with the strongest prediction performance least number of features.

Outline

1. Explore descriptive statistics. Determined suitable pre-processing and machine learning methods.
2. Variables with high missing data (above 40% missing) discarded.
3. Data split into 3 features subset (which are data subsets of the full dataset). Each features subset had separate pipeline and model building.
4. Features subsets were: imputed, oversampled using SMOTENC and then scaled.
5. Modelling of each subset features:
   i) Initial models built.
   ii) Models trained, tested and cross-validated using 5-fold cross validation.
   iii) Prediction performance of models reviewed. Patterns in the model coefficients or important features were compared. Information was used to inform appropriate feature selection strategy.
6. Feature reduction: Genetic Algorithm (GA) feature selection was selected to narrow down features. Features were narrowed to a minimum viable number of features. Built and test models with 2-fold cross validation to determine which reduced model had viable prediction performance.
7. Using the minimum viable number of features, forward SFS (Sequential Feature Selector) was performed to select the best features to fill minimum number. 2-fold cross validation was used to validate the new models.
8. Forward SFS was also used to verify the features selected by the GA feature selection. Forward SFS' features should rank amongst the highest of GA's features.
9. Build model using selected features and select model with best performance, using 5-fold cross validation.
10. Features of the final selected model and prediction performance were reviewed.

GA feature selection (with prediction performance scoring) was picked for its fitness evolution. This allows progressive selection and testing to narrow down features. Forward SFS was picked for its ability to select a predetermined minimum number of features. (Forward SFS was picked over reverse SFS, because the minimum feature number was 5 while the full features close be close to 20.)

Method

2.1) Descriptive Statistics

Statistics were evaluated using the python library 'SciPy' 1.9.1. Continuous data were tested for normality using KS (Kolmogorov–Smirnov) goodness of fit test. Mann-Whitney U test was used to compare continuous variables were compared against intubation outcome. This was because of heavily imbalanced data (97% non-intubated v.s 3% intubated). Non-continuous independent variables were compared against the outcome using the χ^2 contingency table. Gender was categorical. SOFA component scores, and Charlson Comorbidity Index (CCI) were ordinal data. SOFA is multi-component scoring index that measure the stage of organ failure. Each SOFA component measures the stage of failure for each organ system. (Singer, 2016)[1] CCI is a multi-component scoring index used to assess the patient's multi-comorbidity severity. (Charlson et al, 1987)[2]

2.2) Data Processing

Data was from a limited dataset provided for this assignment. Data source was MIMIC-IV (Medical Information Mart for Intensive Care IV), version 2.1. Gender was encoded into binary value (1 = Male, 0 = Female. Initial label order for gender was randomly determined.) Intubation outcome data was analysed as-is in binary (FALSE = not intubated. TRUE = intubated.)

Based on knowledge, the full dataset was *a priori* split into 3 features subsets believed to be useful for interpretation:

- Subsets based on Gender and non-continuous multi-component scores. I.e. SOFA components and Charlson Comorbidity Index (CCI). ('GSC')
- Subsets based on gender and continuous clinical data. ('Gender and continuous data')
- Subsets based on the combination of the above non-continuous and continuous variables. ('Combination')

Intubation is one method of introducing oxygen ICU patients in need of oxygen. All 3 features subsets included gender to determine whether underlying differences in gender physiology might influence the indication or contraindication for intubation.

GSC features subset were conceived based on the value of SOFA components and CCI scores. Literature review indicates that patients might need more oxygen cases upon organ failures and complications severe comorbidities.

Gender and Continuous features subset aims to search for clinical measurements that might predict intubation.

Combination features subset explores combinations of the above variables. This is an alternative all-inclusive approach to model building.

2.3) Missing Data

All 36,489 sample data were used. However, independent variables with high percentage of missing data (> 40% missing) were discarded. This was in line with recommendations for missing medical data. (Jakobsen et al, 2017) [3]

Based on medical knowledge and verification with clinicians, ICU data are often MNAR (Missing Not At Random). Clinicians will only collect specific clinical measurements when they assessed the measurement is needed for diagnosis or monitoring. Hence, the lack of samples only indicate that these are not routine ICU measurements.

2.4) Imputation

For the analysed variables, most of the data had less than 1% missing data. Missing data were assumed to be MAR (Missing At Random) and were imputed. Continuous variables were imputed using KNN (K-Nearest Neighbour) imputation. This was successfully used in a very similar work by Siu et al, 2020 [4]. KNN was shown to achieve good result in various empirical work. However, KNN produced fractional imputes, making it incompatible with ordinal data. Instead, SOFA components and CCI were imputed using the median.

2.5) Oversampling

There was clear imbalance of intubation outcomes (97% non-intubated to 3% intubated). In each pipeline, minority classes were *separately* oversampled using SMOTENC (Synthetic Minority Over-sampling Technique for Nominal and Continuous) from the Python library 'Imbalanced-learn' 0.10.1.

2.6) Scaling

Because Support Vector Machine is sensitive to scaling, scaling was performed. For consistency, scaling was applied across all data in every features subset. This is maintain interoperability when comparing between models.

Every features subset were scaled with either standard or robust scaling. Robust scaling was picked as alternative because of the use of median and interquartile was robust to outliers and the shape of variance distribution. Hence, it was believed that robust scaling might perform better in mixed data types of categorical data (gender), continuous data, and ordinal data.

2.7) Machine Learning

All machine learning methods were performed using the Python library 'Scikit-learn' 1.2.1. K-fold cross validation (K = 5) was used to split each model's data for training, testing and cross validation. Comparison of performance across models were measured using the weighted mean F1 score and AUC ROC (Area Under the Receiver Operating Characteristic Curve). For AUC ROC, OvR (One-vs-the-Rest OvR) multiclass strategy was picked for this study's multiclass design.

The linear model classification used was logistic regression ('Log Reg'). Hyperparameter tuning: C-penalty = 1, 2, 3. Another linear model used was SVM (Support Vector Machine) with SGD (Stochastic Gradient Descent) learning ('SGD SVM'). SVM SGD was used with L2 regularisation. Hyperparameter tuning: alpha = 0.001 to 1.0 (steps = 0.1), fraction of data used to determine early stopping validation 0.1 to 1.0 in steps = 2. Learning rate controlled by alpha, via a heuristic.

Non-linear model classification was performed using C-Support Vector Classification ('SVC rbf') with L2 regularisation. Kernels tested were RBF (Radial Basis Function), polynomial, sigmoid. Gamma determined using modules' default heuristic. Hyperparameter tuning: C-penalty = 1, 2, 3.

For Ensemble learning, out-of-bag bootstrapped Random Forest ('RF') and AdaBoost. Criterion in RF, was entropy. Hyperparameter tuning: number of trees = 50, 100, 150. Learning rate = 0.001 to 1.0 (steps = 0.1). Maximum depth for AdaBoost stump = 1. Hyperparameter tuning: maximum number of stumps to terminate learning = 10, 25, 50. Learning rate = 0.001 to 1.0, steps = 0.1

GA (Genetic Algorithm) feature selection were conducted using the Python library 'Sk-learn-genetic' 0.5.1. Fitness scoring was determined using micro-averaged F1 score. Evolution generations = 22. Fitness and fitness SD (standard deviation) peaked / bottomed well before 22 generations. Another feature selection algorithm used was the forward SFS (Sequential Features Selector) with micro-averaged F1 scoring.

Results

3.1) Descriptive statistics

Table 1 indicates the descriptive statistics of all independent variables used in this study. Initially, the dataset had 59 independent variables and 1 dependent outcome (intubation: TRUE/FALSE). 26 of 59 independent variables exceeded 40% missing data. Remaining 27 of 33 independent variables had missing data between 0 to 1%. 3 of the 33 variables had missing data between 4% to 11% (pt_min = 11.4%, pt_max = 11.4%, urineoutput 3.9%). The last 3 variables (gender, age and CCI) had no missing data.

All continuous variables had normal distribution, as determined by *p*-value (<0.001; $\alpha = 0.05$) for KS (Kolmogorov–Smirnov) goodness of fit to normal distribution. (**Table 2**) Mann Whitney U test against intubation outcomes found age, mbp_mean, dbp_max, temperature_min to be insignificant. (*p*-value marked > α; $\alpha = 0.05$) This suggests the medians for these variables differ across the incubation outcome. These 3 variables might be strong features in the machine learning models. In **Table 3**, the X^2 test of each categorical and ordinal expected incubation outcome were significant. (*p*-value marked > α; $\alpha = 0.05$) This indicates that the observed distribution differs from the expected distribution, suggestion a possible relationship between these variables and the outcome.

There was clear class imbalance of intubation outcomes. 97% of the patients were not intubated while only 4% intubated. (**Table 4**) This finding indicated the need for oversampling methods like SMOTENC. This method produced synthetic minority data points to balance the minority class' impact on the machine learning algorithm's cost function. This was empirically proven on the dataset. Prior to using SMOTE on datasets, various models were tested on the unbalanced dataset. In predicting intubated patients, all the models had F1 score = 0.

3.2) Machine Learning Models

KNN v.s. median imputation

For the same machine learning method and scaling, the two imputation pipelines had hardly any difference in F1 scores and ROC AUC. Any difference in performance was less than 1 percentage point. (**Table 6**) This was true even after hyperparameter training, GA feature selection, and forward SFS. However, for the same machine learning method, imputation method did influence the feature count and which features were selected by GA feature selection.

Standard v.s robust scaling

The same observations for imputations were also observed for standard and robust scaling were compared (**Table 6**).

SVC rbf, SVC polynomial and SVC sigmoid models

Models using GSC features (Gender, SOFA component score, and CCI) had weighted mean F1 and AUC ROC between 0.76 to 0.77. (**Table 6**). The lower prediction performance of GSC features is consistent seen in all other machine learning methods tested in this study (Table 5) Nonetheless, SVC rbf models were very successful in prediction for the other 2 features subsets: (1) Gender and Continuous, and (2) Combination. F1 and AUC ROC scores were consistently 0.98. The high performance was likely due to the RBF kernel matching the cluster pattern of the data points in the hyperplane. Across all features subsets, RBF kernel outperformed the linear SVM SGD in Table 6. Unfortunately, there was not enough computing power to do GA feature selection for RBF kernel model. Further RBF kernel exploration was terminated. The polynomial and sigmoid performed kernels poorly (ROC AUC < 0.70). Thus, kernel modelling was discontinued.

All other models

Interestingly, all other models (after GA feature selection) had similar prediction performance in Gender and Continuous features subset and Combination subset. Weighted mean F1 and ROC AUC were 0.97 ± across all models after GA feature selection. (**Table 5**)

For all models (after GA feature selection), the prediction performance in GSC did not improve much beyond their F1 and ROC AUC scores prior to GA feature selection. (**Table 5**) Random Forest had better performance (about ROC AUC = 0.85), above the other models (0.69 to 0.74).

Post GA selection, the most important features were mapped across all models for comparison (**Figure 2**). Approximately 5 to 7 continuous features are features with significant *p*-value for the Mann Whitney U test.

Following some experimentation, the minimum viable feature count was 5 for Gender and Continuous features subset. This was confirmed after evaluating the prediction performance of the models after forward SFS picked the 5 best features. (As compared to 6 or 7.)

Following SFS, the best model among all models was Random Forest, with features: age, dbp_mean, temperature_mean, wbc-max and creatinine min. (Imputation and scaling did not affect the model.) This model had AUC ROC of 0.98 and the least number of features. Other models saw drop in AUC ROC to around 0.93 to 0.96 following feature selection.

Figure 1 shows AUC ROC curve of 3 of the best performing median impute, robust scaling models. (Random Forest AUC ROC = 0.98, AdaBoost 0.97, Logistic Regression 0.96).

Discussion

SVC with RBF kernel was shown to outperform linear SVM method in this study, suggesting that linear SVM is unlikely to yield promising result.

This also suggest that the ideal model in the hyperplane is not linear. This also explains the poor performance of logistic regression.

This leaves ensemble methods that are technically scale and distribution insensitive as promising model. In this study, Random Forest outperformed AdaBoost. AdaBoost is a more accurate model when weak classifiers are important features. However, in this dataset, the emphasis on weak classifiers might had contributed to the larger number of features required to yield a high AUC ROC (see **Figure 2**). Random Forest might have performed better, because of the underlying physiology influencing intubation. Physiology is often influenced by multiple variables at play simultaneously. For such a situation, Random Forest does

better than AdaBoost. Unlike AdaBoost's emphasis on weak classifiers, Random Forest employs equal vote among the forest of decision trees.

For this study, in terms of pragmatics, Random Forest avoided features with missing percentage of 3.9% to 11.3%. This makes the Random Forest model (KNN / median impute, standard / robust scaling) to be the most power predictor, with the least features, and using features that are almost certainly measured for all ICU patients.

References

[1] Singer, Mervyn; et al. (23 February 2016). "The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)". JAMA. 315 (8): 801–10. doi:10.1001/jama.2016.0287

[2] Charlson, Mary E.; Pompei, Peter; Ales, Kathy L.; MacKenzie, C. Ronald (1987). "A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation". Journal of Chronic Diseases. 40 (5): 373–83. doi:10.1016/0021-9681(87)90171-8.

[3] Jakobsen, Janus Christian, et al. "When and how should multiple imputation be used for handling missing data in randomised clinical trials–a practical guide with flowcharts." BMC medical research methodology 17.1 (2017): 1-10.

[4] Siu, Benjamin Ming Kit, et al. "Predicting the need for intubation in the first 24 h after critical care admission using machine learning approaches." Scientific reports 10.1 (2020): 20931.

Figures and Tables

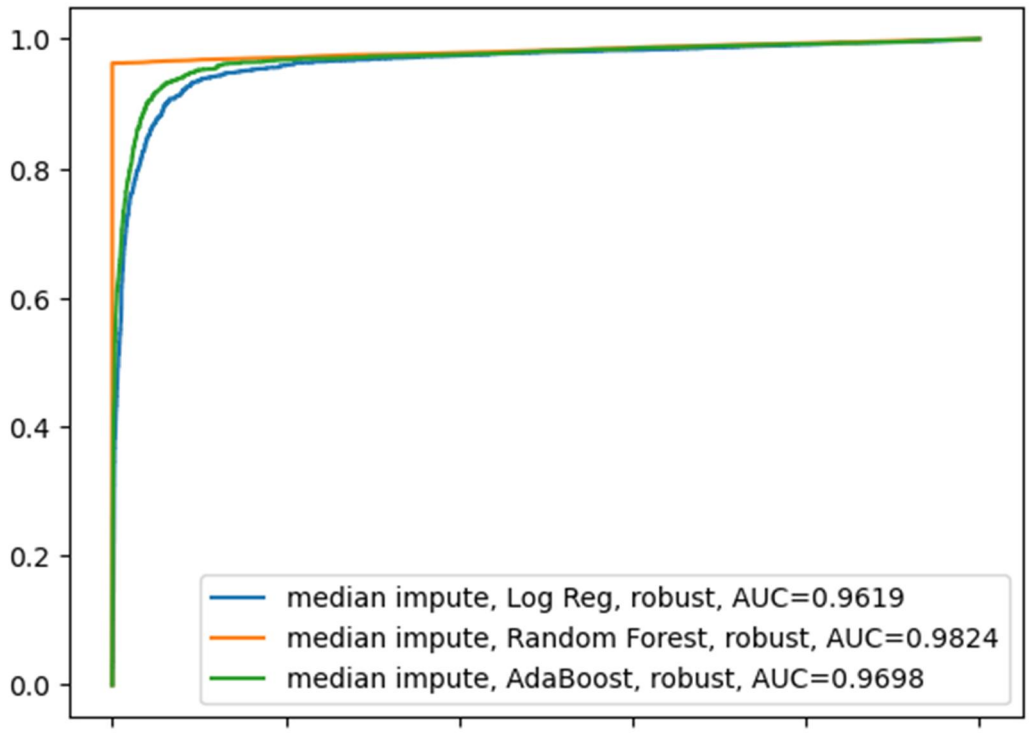


Figure 1: AUC ROC curve for 3 selected best models (feature count of 5, after SFS)

Table 1: (Below) Descriptive statistics of analysed data. SD = standard deviation. Yellow rows highlight variables that had more missing data that stood out among the analysed data (missing % between 3.9 to 11.4%). N = 36,489 patients

	Count	Mean	SD	Min	25th quartile	Median	75th quartile	Max	Missing %
age	36,489	65.9	16.8	18	55.5	67.4	78.7	102.9	0.0
heart_rate_min	36,417	70.9	15.1	9	60	70	80	163	0.2
heart_rate_max	36,417	103.1	20.7	36	88	101	115	295	0.2
heart_rate_mean	36,417	84.9	16	28.5	73.3	83.5	95.3	174.7	0.2
mbp_min	36,401	60.4	13.3	0.8	53	60	68	133	0.2
mbp_max	36,401	105.5	23.3	51	91	102	115	299	0.2
mbp_mean	36,401	79.3	11.8	42.5	70.9	77.9	86.4	151.5	0.2
sbp_min	36,343	94.4	16.7	2	83.5	93	104	184	0.4
sbp_max	36,343	147.7	22.9	49	132	146	161	352	0.4
sbp_mean	36,343	119.8	17.2	40	107.1	117.7	130.6	206.4	0.4
dbp_min	36,341	47.4	11.9	1	40	47	54	113	0.4
dbp_max	36,341	90.4	20.2	29	76	88	101	273	0.4
dbp_mean	36,341	64.5	11.7	29	56.3	63.3	71.6	130.4	0.4
temperature_min	36,126	36.3	0.5	22	36.2	36.4	36.7	38.6	1.0
temperature_max	36,126	37.3	0.6	33.4	36.9	37.2	37.6	42	1.0
temperature_mean	36,126	36.8	0.4	33.2	36.6	36.8	37	39.5	1.0
glucose_min	36,191	117.7	39.9	7	93	109	132	575	0.8
glucose_max	36,191	167.3	113.2	7	113	137	178	2,440.00	0.8
wbc_min	36,235	10.2	7.8	0.1	6.6	9	12.2	300.4	0.7
wbc_max	36,235	13.1	11.1	0.1	8.3	11.4	15.7	407.2	0.7
creatinine_min	36,291	1.3	1.5	0.1	0.7	0.9	1.3	31.8	0.5
creatinine_max	36,291	1.6	1.8	0.1	0.8	1	1.6	43	0.5
hemoglobin_min	36,232	10.3	2.3	2.2	8.6	10.3	11.9	19.3	0.7
hemoglobin_max	36,232	11.4	2.2	3.7	9.8	11.3	12.9	21.7	0.7
pt_min	32,345	14.9	6.6	8	11.9	13.1	15.1	140.3	11.4
pt_max	32,345	17.1	11.3	8	12.3	13.9	16.8	154.6	11.4
urineoutput	35,073	1,915	1,267	-14,850	1,050	1,667	2,500	31,016	3.9
sofa_coagulation	36,235	N.a.	N.a.	0	0	0	1	4	0.7
sofa_cardiovascular	36,402	N.a.	N.a.	0	1	1	1	4	0.2
sofa_cns	36,468	N.a.	N.a.	0	0	1	1	4	0.1
sofa_renal	36,467	N.a.	N.a.	0	0	0	1	4	0.1
CCI	36,489	N.a.	N.a.	0	4	6	8	20	0.0
gender	36,489	N.a.	N.a.	N.a.	N.a.	N.a.	N.a.	N.a.	0.0
outcome	36,489	N.a.	N.a.	N.a.	N.a.	N.a.	N.a.	N.a.	0.0

Table 2: (Right) Statistic test on continuous variables. p-values for KS (Kolmogorov–Smirnov) test for normality and Mann Whitney U test against intubation outcomes are indicated. p-value marked (*) are significant (α = 0.05). Yellow highlights indicate insignificant p-values.

	KS test p-value	Mann Whitney U p-value
age	< 0.001*	0.19
heart_rate_min	< 0.001*	< 0.001*
heart_rate_max	< 0.001*	< 0.001*
heart_rate_mean	< 0.001*	< 0.001*
mbp_min	< 0.001*	< 0.001*
mbp_max	< 0.001*	0.04*
mbp_mean	< 0.001*	0.07
sbp_min	< 0.001*	< 0.001*
sbp_max	< 0.001*	0.006*
sbp_mean	< 0.001*	0.025*
dbp_min	< 0.001*	< 0.001*
dbp_max	< 0.001*	0.69
dbp_mean	< 0.001*	0.014*
temperature_min	< 0.001*	0.33
temperature_max	< 0.001*	< 0.001*
temperature_mean	< 0.001*	< 0.001*
glucose_min	< 0.001*	0.008*
glucose_max	< 0.001*	< 0.001*
wbc_min	< 0.001*	< 0.001*
wbc_max	< 0.001*	< 0.001*
creatinine_min	< 0.001*	< 0.001*
creatinine_max	< 0.001*	< 0.001*
hemoglobin_min	< 0.001*	< 0.001*
hemoglobin_max	< 0.001*	0.025*
pt_min	< 0.001*	< 0.001*
pt_max	< 0.001*	< 0.001*
urineoutput	< 0.001*	< 0.001*

Table 6: (Right) Summary of SVC rbf models. GSC = Gender, SOFA component score, and CCI. SVC rbf = C-Support Vector Classification with RBF (Radial Basis Function) kernel

Table 5: (Right) Summary of all other models built and tested. GSC = Gender, SOFA component score, and CCI. Log Reg = Logistic Regression. GA = Genetic Algorithm (GA) feature selection. SFS = forward Sequential Features Selector.

Table 3: (Bottom) X^2 test of each categorical and ordinal expected incubation outcome. p-value marked (*) are significant (α = 0.05).

	X^2 test p-value
gender	< 0.001*
sofa_coagulation	< 0.001*
sofa_cardiovascular	< 0.001*
sofa_cns	< 0.001*
sofa_renal	< 0.001*
CCI	0.002*

Table 4: (Bottom) Breakdown of intubation outcomes across gender count.

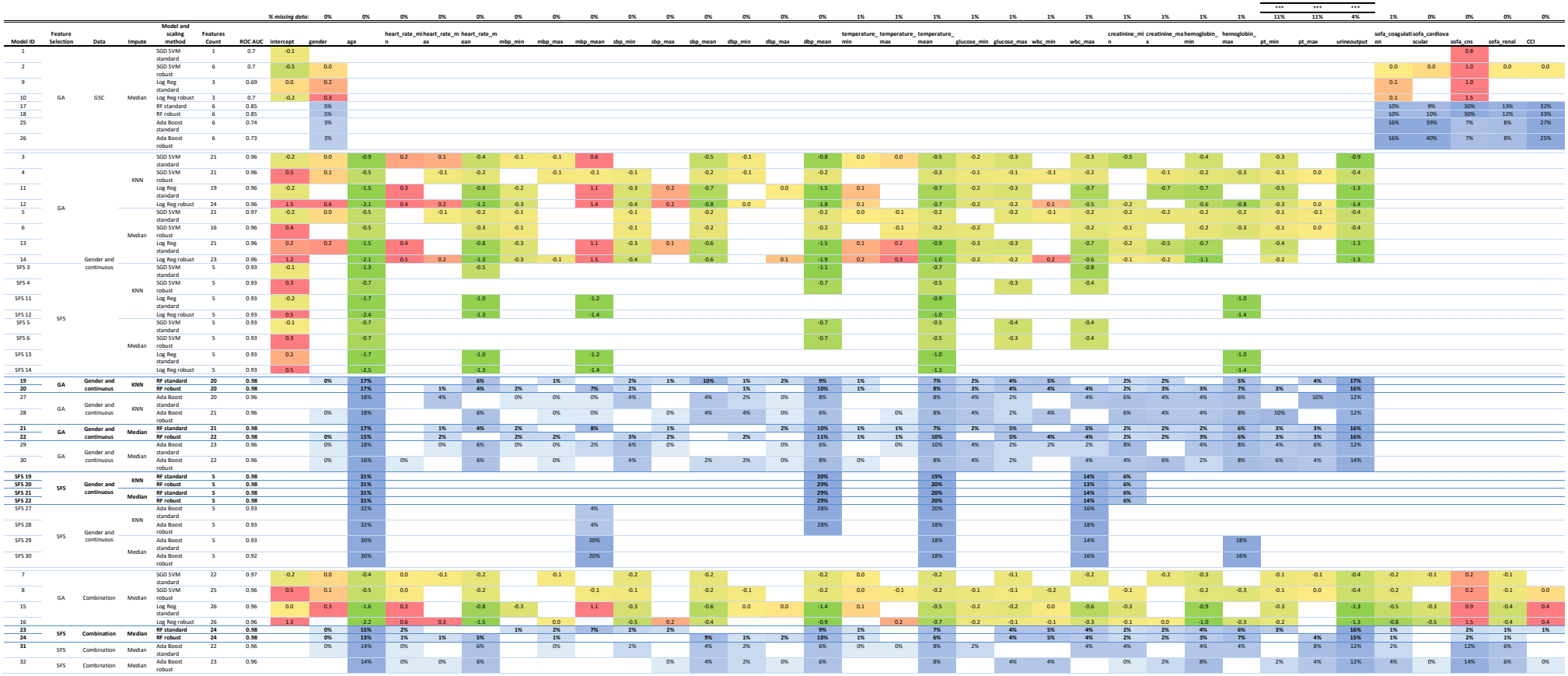
	Not intubated	Intubated	Total
F	16,079	530	**16,609**
M	19,081	799	**19,880**
Total	**35,160**	**1,329**	

	Not intubated	Intubated	Total
F	97%	3%	**100%**
M	96%	4%	**100%**

	Not intubated	Intubated
F	46%	40%
M	54%	60%
Total	**100%**	**100%**

Model ID	Feature Selection	Data	Impute	Model and scaling method	Intubation False F1 score	Intubation True F1 score	Weighted mean F1 Score	ROC AUC OvR
A	None	GSC	Median	SVC rbf standard	0.77	0.76	0.77	0.77
B				SVC rbf robust	0.77	0.76	0.76	0.76
C	None	Gender and continuous metrics	KNN	SVC rbf standard	0.98	0.98	0.98	0.98
D				SVC rbf robust	0.98	0.98	0.98	0.98
E	None	Gender and continuous metrics	Median	SVC rbf standard	0.98	0.98	0.98	0.98
F				SVC rbf robust	0.98	0.98	0.98	0.98
G	None	Combination	Median	SVC rbf standard	0.98	0.98	0.98	0.98
H				SVC rbf robust	0.98	0.98	0.98	0.98

Model ID	Feature Selection	Data	Impute	Model and scaling method	Intubation False F1 score	Intubation True F1 score	Weighted mean F1 Score	ROC AUC OvR
1	GA	GSC	Median	SGD SVM standard	0.73	0.66	0.69	0.70
2				SGD SVM robust	0.74	0.65	0.69	0.70
9				Log Reg standard	0.72	0.66	0.69	0.69
10				Log Reg robust	0.73	0.66	0.69	0.70
17				Random Forest standard	0.84	0.85	0.85	0.85
18				Random Forest robust	0.84	0.85	0.84	0.85
25				Ada Boost standard	0.74	0.73	0.74	0.74
26				Ada Boost robust	0.74	0.73	0.73	0.73
3	GA	Gender and continuous metrics	KNN	SGD SVM standard	0.96	0.96	0.96	0.96
4				SGD SVM robust	0.96	0.96	0.96	0.96
11				Log Reg standard	0.96	0.96	0.96	0.96
12				Log Reg robust	0.96	0.96	0.96	0.96
19				Random Forest standard	0.98	0.98	0.98	0.98
20				Random Forest robust	0.98	0.98	0.98	0.98
27				Ada Boost standard	0.96	0.96	0.96	0.96
28				Ada Boost robust	0.96	0.96	0.96	0.96
5	GA	Gender and continuous metrics	Median	SGD SVM standard	0.96	0.96	0.96	0.97
6				SGD SVM robust	0.96	0.96	0.96	0.96
13				Log Reg standard	0.96	0.96	0.96	0.96
14				Log Reg robust	0.96	0.96	0.96	0.96
21				Random Forest standard	0.98	0.98	0.98	0.98
22				Random Forest robust	0.98	0.98	0.98	0.98
29				Ada Boost standard	0.96	0.96	0.96	0.96
30				Ada Boost robust	0.96	0.96	0.96	0.96
7	GA	Combination	Median	SGD SVM standard	0.97	0.97	0.97	0.97
8				SGD SVM robust	0.96	0.96	0.96	0.96
15				Log Reg standard	0.97	0.96	0.96	0.96
16				Log Reg robust	0.96	0.96	0.96	0.96
23				Random Forest standard	0.98	0.98	0.98	0.98
24				Random Forest robust	0.98	0.98	0.98	0.98
31				Ada Boost standard	0.96	0.96	0.96	0.96
32				Ada Boost robust	0.96	0.96	0.96	0.96
SFS 3	SFS	Gender and continuous metrics	KNN	SGD SVM standard	0.93	0.93	0.93	0.93
SFS 4				SGD SVM robust	0.93	0.93	0.93	0.93
SFS 11				Log Reg standard	0.93	0.93	0.93	0.93
SFS 12				Log Reg robust	0.93	0.93	0.93	0.93
SFS 19	SFS	Gender and continuous metrics	KNN	Random Forest standard	0.98	0.98	0.98	0.98
SFS 20				Random Forest robust	0.98	0.98	0.98	0.98
SFS 27	SFS	Gender and continuous metrics	KNN	Ada Boost standard	0.93	0.93	0.93	0.93
SFS 28				Ada Boost robust	0.93	0.93	0.93	0.93
SFS 5	SFS	Gender and continuous metrics	Median	SGD SVM standard	0.93	0.93	0.93	0.93
SFS 6				SGD SVM robust	0.93	0.93	0.93	0.93
SFS 13				Log Reg standard	0.93	0.93	0.93	0.93
SFS 14				Log Reg robust	0.93	0.93	0.93	0.93
SFS 21	SFS	Gender and continuous metrics	Median	Random Forest standard	0.98	0.98	0.98	0.98
SFS 22				Random Forest robust	0.98	0.98	0.98	0.98
SFS 29	SFS	Gender and continuous metrics	Median	Ada Boost standard	0.93	0.93	0.93	0.93
SFS 30				Ada Boost robust	0.92	0.92	0.92	0.92



Model ID	Feature Selection	Data	Impute	Model and scaling method	Features Count	ROC AUC	intercept	gender	age	heart_rate_min	heart_rate_max	heart_rate_mean	mbp_min	mbp_max	mbp_mean	sbp_min	sbp_max	sbp_mean	dbp_min	dbp_max	dbp_mean	temperature_min	temperature_max	temperature_mean	glucose_min	glucose_max	wbc_min	wbc_max	creatinine_min	creatinine_max	hemoglobin_min	hemoglobin_max	pt_min	pt_max	urineoutput	sofa_coagulation	sofa_cardiovascular	sofa_cns	sofa_renal	CCI
				% missing data:			0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	*** 11%	*** 11%	*** 4%	1%	0%	0%	0%	0%
1				SGD SVM standard	1	0.7	-0.1																														0.8		
2				SGD SVM robust	6	0.7	-0.5	0.0																											0.0	0.0	1.0	0.0	0.0
9				Log Reg standard	3	0.69	0.0	0.2																											0.1		1.0		
10	GA	GSC	Median	Log Reg robust	3	0.7	-0.2	0.3																											0.1		1.5		
17				RF standard	6	0.85		5%																											10%	9%	30%	13%	32%
18				RF robust	6	0.85		5%																											10%	10%	30%	12%	33%
25				Ada Boost standard	6	0.74		3%																											16%	39%	7%	8%	27%
26				Ada Boost robust	6	0.73		3%																											16%	40%	7%	8%	25%
3				SGD SVM standard	21	0.96	-0.2	0.0	-0.9	0.2	0.1	-0.4	-0.1	-0.1	0.6			-0.5	-0.1		-0.8	0.0	0.0	-0.5	-0.2	-0.3		-0.3	-0.5		-0.4		-0.3		-0.9					
4			KNN	SGD SVM robust	21	0.96	0.5	0.1	-0.5		-0.1	-0.2		-0.1	-0.1	-0.1		-0.2	-0.1		-0.2			-0.3	-0.1	-0.1	-0.1	-0.2		-0.1	-0.2	-0.3	-0.1	0.0	-0.4					
11				Log Reg standard	19	0.96	-0.2		-1.5	0.3		-0.8	-0.2		1.1	-0.3	0.2	-0.7		0.0	-1.5	0.1		-0.7	-0.2	-0.3		-0.7		-0.7	-0.7		-0.5		-1.3					
12				Log Reg robust	24	0.96	1.5	0.6	-2.1	0.4	0.2	-1.2	-0.3		1.4	-0.4	0.2	-0.8	0.0		-1.8	0.1		-0.7	-0.2	-0.2	0.1	-0.5	-0.2		-0.6	-0.8	-0.3	0.0	-1.4					
5	GA			SGD SVM standard	21	0.97	-0.2	0.0	-0.5		-0.1	-0.2	-0.1			-0.1		-0.2			-0.2	0.0	-0.1	-0.2		-0.2	-0.1	-0.2	-0.2	-0.2	-0.2	-0.2	-0.1	-0.1	-0.4					
6			Median	SGD SVM robust	16	0.96	0.4		-0.5			-0.3	-0.1			-0.1		-0.2			-0.2		-0.1	-0.2	-0.2				-0.2	-0.1		-0.2	-0.3	-0.1	0.0	-0.4					
13				Log Reg standard	21	0.96	0.2	0.2	-1.5	0.4		-0.8	-0.3		1.1	-0.3	0.1	-0.6			-1.5	0.1	0.2	-0.9	-0.3	-0.3		-0.7	-0.2	-0.5	-0.7		-0.4		-1.3					
14		Gender and continuous		Log Reg robust	23	0.96	1.2		-2.1	0.5	0.2	-1.3	-0.3	-0.1	1.5	-0.4		-0.6		0.1	-1.9	0.2	0.3	-1.0	-0.2	-0.2	0.2	-0.6	-0.1	-0.2	-1.1		-0.2		-1.3					
SFS 3				SGD SVM standard	5	0.93	-0.1		-1.3			-0.5									-1.1			-0.7				-0.8											
SFS 4			KNN	SGD SVM robust	5	0.93	0.3		-0.7												-0.7			-0.5		-0.3		-0.4											
SFS 11				Log Reg standard	5	0.93	-0.2		-1.7			-1.0			-1.2									-0.9							-1.0								
SFS 12	SFS			Log Reg robust	5	0.93	0.5		-2.4			-1.3			-1.4									-1.0							-1.4								
SFS 5				SGD SVM standard	5	0.93	-0.1		-0.7												-0.7			-0.5		-0.4		-0.4											
SFS 6			Median	SGD SVM robust	5	0.93	0.3		-0.7												-0.7			-0.5		-0.3		-0.4											
SFS 13				Log Reg standard	5	0.93	0.2		-1.7			-1.0			-1.2									-1.0							-1.0								
SFS 14				Log Reg robust	5	0.93	0.5		-2.5			-1.3			-1.4									-1.1							-1.4								
19	**GA**	**Gender and continuous**	**KNN**	**RF standard**	**20**	**0.98**		**0%**	**17%**			**6%**		**1%**		**2%**	**1%**	**10%**	**1%**	**2%**	**9%**	**1%**		**7%**	**2%**	**4%**	**5%**		**2%**	**2%**		**5%**		**4%**	**17%**					
20				**RF robust**	**20**	**0.98**			**17%**		**1%**	**4%**	**2%**		**7%**	**2%**			**1%**		**10%**	**1%**		**8%**	**3%**	**4%**	**4%**	**4%**	**2%**	**3%**	**3%**	**7%**	**3%**		**16%**					
27				Ada Boost standard	20	0.96			18%		4%		0%	0%	0%	4%		4%	2%	0%	8%			8%	4%	2%		4%	6%	4%	4%	6%		10%	12%					
28	GA	Gender and continuous	KNN	Ada Boost robust	21	0.96		0%	18%			6%		0%	0%		0%	4%	4%	0%	6%		0%	8%	4%	2%	4%		6%	4%	4%	8%	10%		12%					
21	**GA**	**Gender and continuous**	**Median**	**RF standard**	**21**	**0.98**			**17%**		**1%**	**4%**	**2%**		**8%**		**1%**			**2%**	**10%**	**1%**	**1%**	**7%**	**2%**	**5%**		**5%**	**2%**	**2%**	**6%**	**3%**	**3%**	**16%**					
22				**RF robust**	**22**	**0.98**		**0%**	**15%**		**2%**		**2%**	**2%**		**3%**	**2%**		**2%**		**11%**	**1%**	**1%**	**10%**		**5%**	**4%**	**4%**	**2%**	**2%**	**3%**	**6%**	**3%**	**3%**	**16%**					
29	GA	Gender and continuous	Median	Ada Boost standard	23	0.96		0%	18%		0%	6%	0%	0%	2%	6%	0%			0%	6%		0%	10%	4%	2%	2%	2%	8%		4%	8%	4%	6%	12%					
30				Ada Boost robust	22	0.96		0%	16%	0%		6%		0%		4%		2%	2%	0%	8%	0%		8%	4%	2%		4%	4%	6%	2%	8%	6%	4%	14%					
SFS 19			**KNN**	**RF standard**	**5**	**0.98**			**31%**												**30%**			**19%**				**14%**	**6%**										
SFS 20	**SFS**	**Gender and continuous**		**RF robust**	**5**	**0.98**			**31%**												**29%**			**20%**				**13%**	**6%**										
SFS 21			**Median**	**RF standard**	**5**	**0.98**			**31%**												**29%**			**20%**				**14%**	**6%**										
SFS 22				**RF robust**	**5**	**0.98**			**31%**												**29%**			**20%**				**14%**	**6%**										
SFS 27			KNN	Ada Boost standard	5	0.93			32%						4%						28%			20%				16%											
SFS 28	SFS	Gender and continuous		Ada Boost robust	5	0.93			32%						4%						28%			18%				18%											
SFS 29			Median	Ada Boost standard	5	0.93			30%						20%									18%				14%			18%								
SFS 30				Ada Boost robust	5	0.92			30%						20%									18%				16%			16%								
7				SGD SVM standard	22	0.97	-0.2	0.0	-0.4	0.0	-0.1	-0.2		-0.1		-0.2		-0.2			-0.2	0.0		-0.2		-0.1		-0.2		-0.2	-0.3		-0.1	-0.1	-0.4	-0.2	-0.1	0.2	-0.1	
8	GA	Combination	Median	SGD SVM robust	25	0.96	0.5	0.1	-0.5	0.0		-0.2			-0.1	-0.1		-0.2	-0.1		-0.2	0.0	-0.1	-0.2	-0.1	-0.1	-0.2		-0.1		-0.2	-0.3	-0.1	0.0	-0.4	-0.2		0.2	-0.1	0.0
15				Log Reg standard	26	0.96	0.0	0.3	-1.6	0.3		-0.8	-0.3		1.1	-0.3		-0.6	0.0	0.0	-1.4	0.1		-0.5	-0.2	-0.2	0.0	-0.6	-0.3			-0.9		-0.3		-1.3	-0.5	-0.3	0.9	-0.4	0.4
16				Log Reg robust	26	0.96	1.3		-2.2	0.6	0.3	-1.5		0.0		-0.5	0.2	-0.4			-0.9		0.2	-0.7	-0.2	-0.1	-0.1	-0.3	-0.1	0.0	-1.0	-0.3	-0.2		-1.3	-0.8	-0.5	1.5	-0.4	0.4
23				**RF standard**	**24**	**0.98**		**0%**	**15%**	**2%**			**1%**	**2%**	**7%**	**2%**	**2%**				**9%**	**1%**		**7%**		**4%**	**5%**	**4%**	**2%**	**2%**	**4%**	**6%**	**3%**		**16%**	**1%**		**2%**	**1%**	**1%**
24	**SFS**	**Combination**	**Median**	**RF robust**	**24**	**0.98**		**0%**	**13%**	**1%**	**1%**	**5%**		**1%**				**9%**	**1%**	**2%**	**10%**	**1%**		**6%**		**4%**	**5%**	**4%**	**2%**	**2%**	**3%**	**7%**		**4%**	**15%**	**1%**		**2%**	**1%**	
31	SFS	Combination	Median	Ada Boost standard	22	0.96		0%	14%	0%		6%		0%		2%		4%	2%		6%	0%	0%	8%	2%			4%	4%		4%	4%		8%	12%	2%		12%	6%	
32	SFS	Combination	Median	Ada Boost robust	23	0.96			14%	0%	0%	6%					0%	4%	2%	0%	6%			8%		4%	4%		0%	2%	8%		2%	4%	12%	4%	0%	14%	6%	0%

Figure 2: (Above) Heat map summarising the various model coefficients, intercept and feature importance. Bold text indicate models with the best ROC AUC scores (ROC AUC = 0.98). Variables marked (*) indicate variables with missing values beyond 1%. For analysed variables had missing values imputed at the start of the pipeline. Models with GA feature selection only had features selected from this algorithm. Models with (forward) SFS selected features for a minimum feature number, that was informed by the GA feature selection.

Legend: Linear models -> Graded from red to yellow to green (0, 50th, and 90th percentile)
Ensemble models -> Graded in increasing darkness of blue, by ascending percentile.