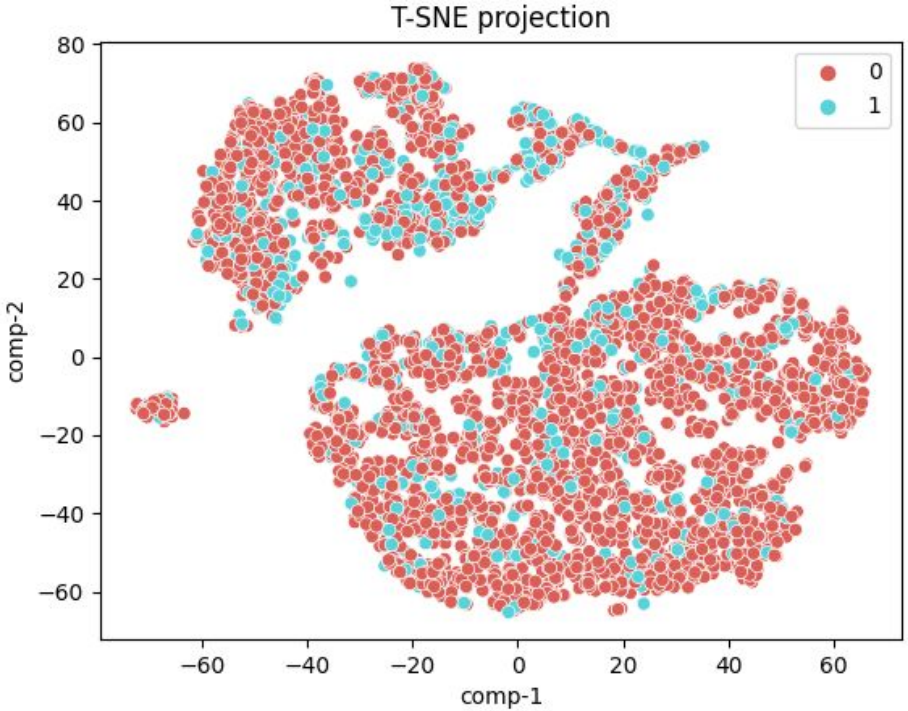
Predicting Sepsis using various machine learning algorithms

Yannick Müller

T-SNE Map


T-SNE projection
0
1
comp-2
comp-1
80
60
40
20
0
−20
−40
−60
−60
−40
−20
0
20
40
60

Preprocessing

1. Sorted by Patient ID so that no two patients can be in two different sets
2. Factorized all text entries in the dataset
3. Deleted rows where the target value was NaN
4. Filled missing values with 0
5. Created new features using the 0/1 Encoding for different percentiles

What does not work: Standardization, Normalization, Feature Operations

Minimal Example

```
25 kf = KFold(n_splits=5)
26 roc_auc_list = []
27 roc_auc_list_int = []
28 f1_list = []
29
30 for train_index, test_index in kf.split(X, y):
31         X_train, X_test = X.iloc[train_index], X.iloc[test_index]
32         y_train, y_test = y.iloc[train_index], y.iloc[test_index]
33         y_flatten = y_train.values.flatten()
34         bin_count = np.bincount(y_flatten)
35         class_weight = bin_count[0]/bin_count[1] - 1
36         clf = Ridge()
37         clf.fit(X_train, y_train, sample_weight=class_weight*y_flatten+1)
38         y_pred = clf.predict(X_test)
39         y_pred_int = list(map(minmaxint, y_pred))
40         roc_auc_list.append(roc_auc_score(y_test, y_pred))
41         roc_auc_list_int.append(roc_auc_score(y_test, y_pred_int))
42         f1_list.append(f1_score(y_test, y_pred_int))
43
44 print("This model:  ", np.mean(roc_auc_list), np.mean(roc_auc_list_int), np.mean(f1_list))
45 print("Current Best: 0.7183366311037686 0.6764051644236325 0.45046285449129")
```

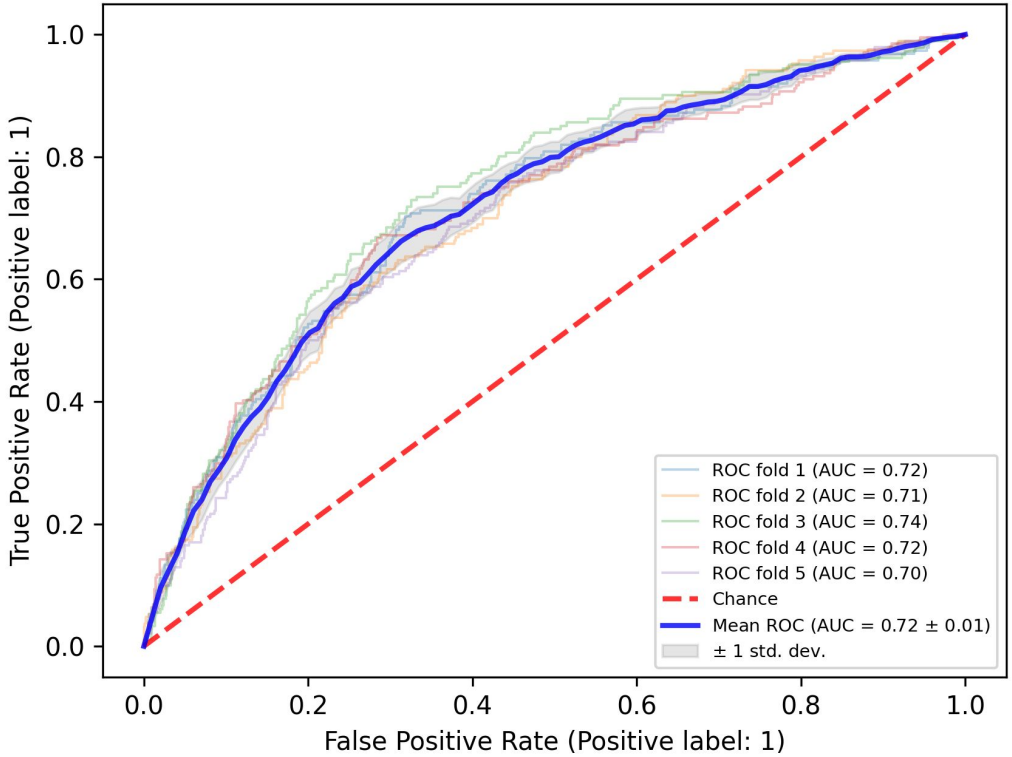
Features

Lactat	0.166	Respir. r. f. 30% 0/1 Enc	0.100	EOS	-0.045
Lactat 70% - 0/1 Enc	0.165	THZ 30% 0/1 Enc	-0.097	Diastolic Bp first	-0.041
Harnstoff 30% - 0/1 Enc	0.164	Lvl consc alert	0.088	Systolic Bp 40% 0/1 Enc	-0.041
INRiH 50 - 0/1 Enc	0.155	THZ	-0.088	Spo2 first	-0.035
CRP	0.142	Hb	-0.075	**Triage**	**-0.034**
BIC_st	0.138	Leuk inf 0.5	0.065	Temp Low 90% 0/1 Enc	0.034
ASAT 30% - 0/1 Enc	0.129	Leuk sup 10	0.060	Gcs inf first	0.024
Zuhause/Altersheim/etc	**0.128**	EOS 60% 0/1 Enc	-0.059	**Inselklinik**	**-0.015**
GGT	0.126	EOS 80% 0/1 Enc	-0.053	Kalium	0.005
Age	0.110	Leuk inf 4	0.050	Referral	-0.003

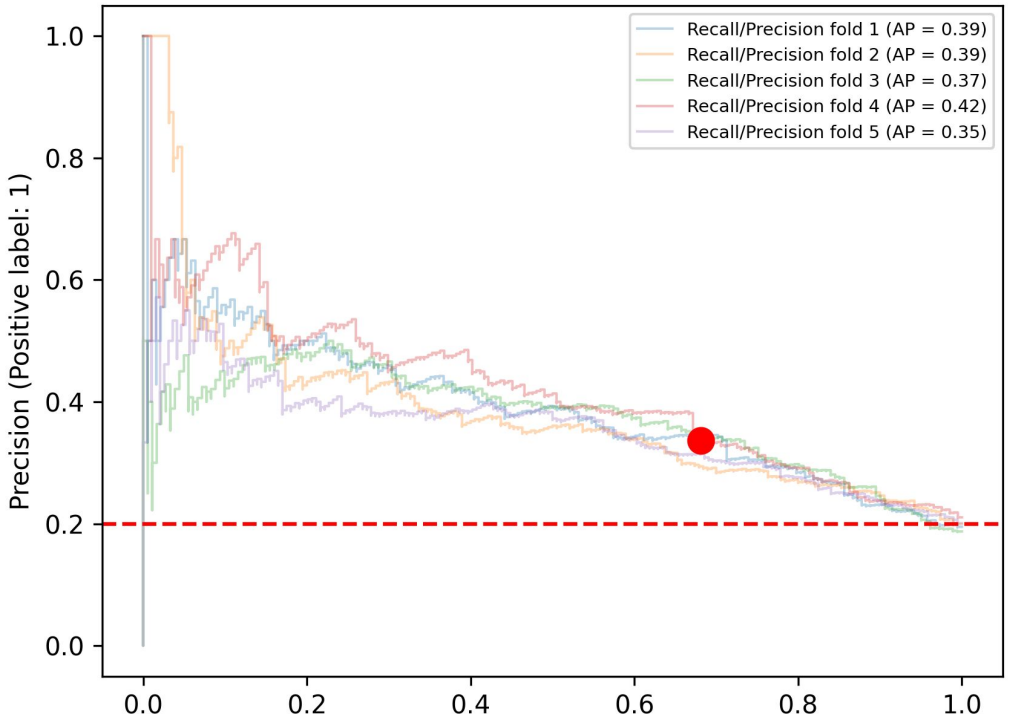
Results

Method	AUC ROC	F1 Score	Recall	Precision	Accuracy
Linear Regression	0.72	0.45	0.68	0.34	0.67
Ridge Regression	0.72	0.45	0.68	0.34	0.67
Dense NN 1 Neuron	0.72	0.44	0.68	0.33	0.67
XGBoost Regressor	0.68	0.41	0.51	0.34	0.71
Random Forest	0.68	0.14	0.08	0.50	0.80
Gaussian NB	0.65	0.43	0.56	0.34	0.70
SVC RBF	0.64	0.41	0.56	0.32	0.68
Lasso	0.67	0.40	0.58	0.31	0.67
MLP Classifier	0.54	0.29	0.37	0.26	0.63
Decision Tree	0.54	0.26	0.26	0.26	0.70

AUC ROC Curve



Recall / Precision Curve



With emojis

Questions

1. Combining the data of different time samples of the same patient improved the model just a bit, dismiss it?
2. Is 0.72 ROC_AUC score high enough to be helpful?
3. Papers that perform better have more than twice as many patients and more features, specifically to predict sepsis. Can we enrich this dataset?
4. In papers, so many things seem to work, like Ensemble learning. Am I doing something wrong, or is it the dataset?
5. In papers, they predict whether the patient has sepsis in four hours. Would that kind of prediction be helpful for you?