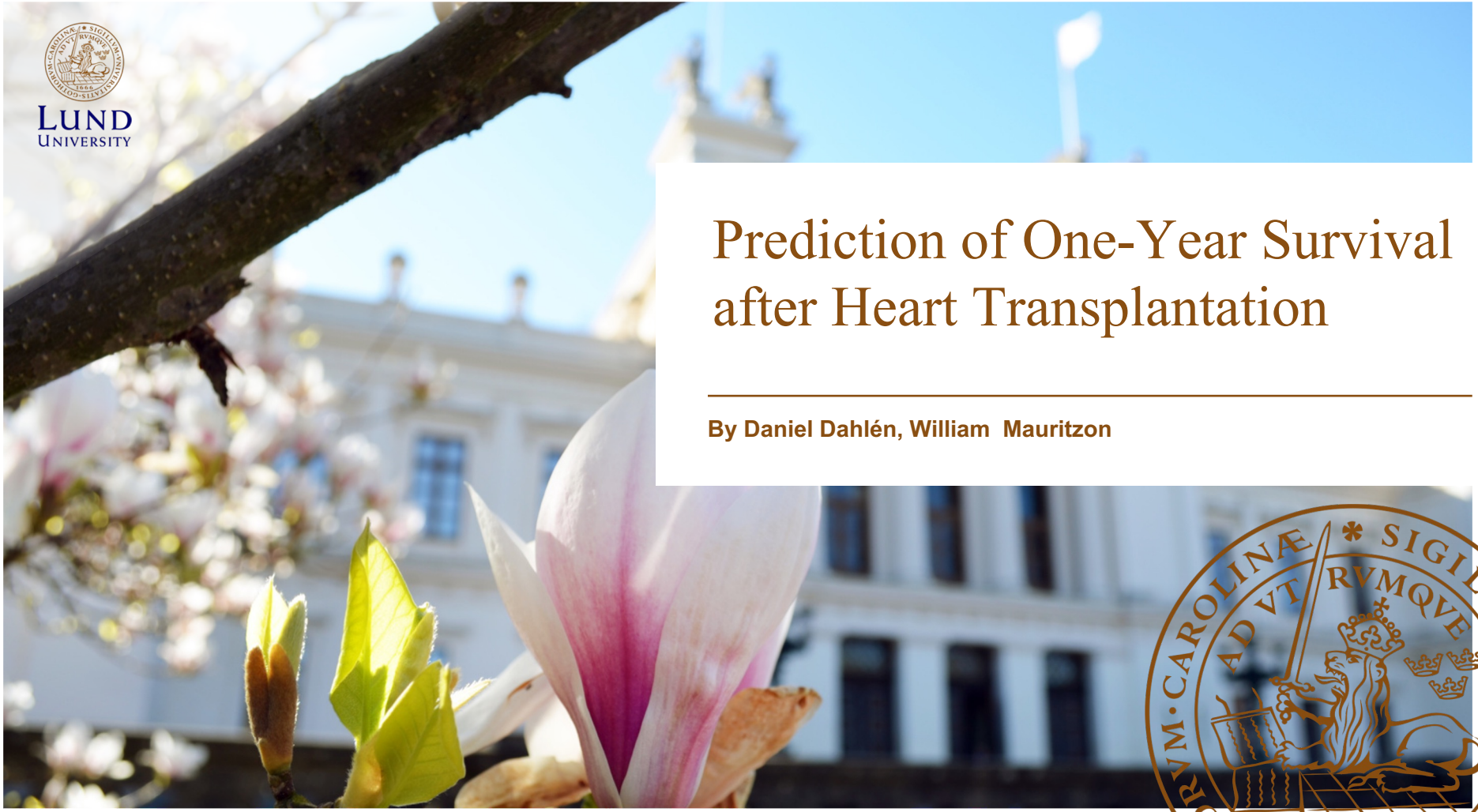
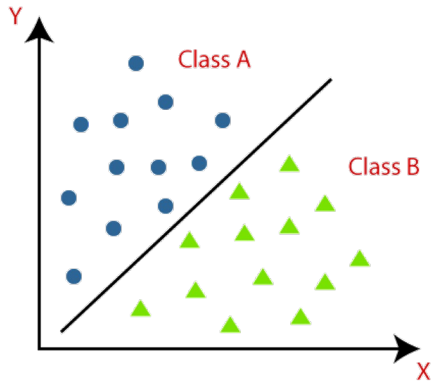


Prediction of One-Year Survival after Heart Transplantation

By Daniel Dahlén, William Mauritzon

The classification problem

Class 0	**Class 1**
The patient will not live longer than one year	The patient will live longer than one year

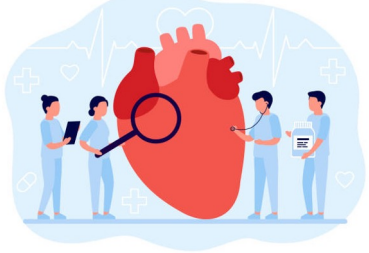


Purpose of project

- Predict the classes based on patient data

- Identify some key features that contribute to mortality outcomes

Data set

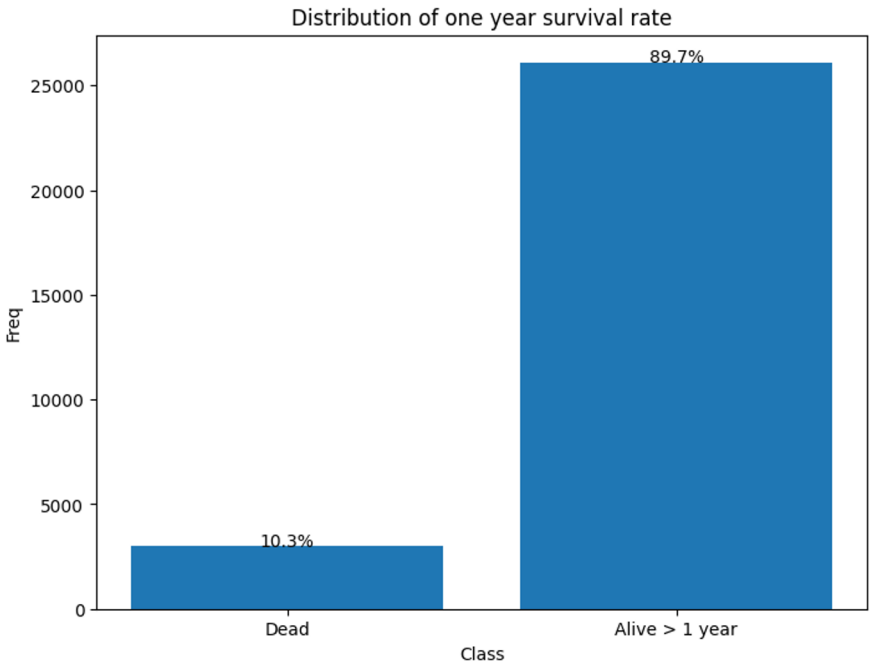


- The data set of heart transplant patients was gathered from the UNOS database (United Network for Organ Sharing)
- Data from 1987
- 500 variables including patient, donor, and transplant information.
- 120 000 patients
- Selected only Heart transplant patients
- 66 000 patients

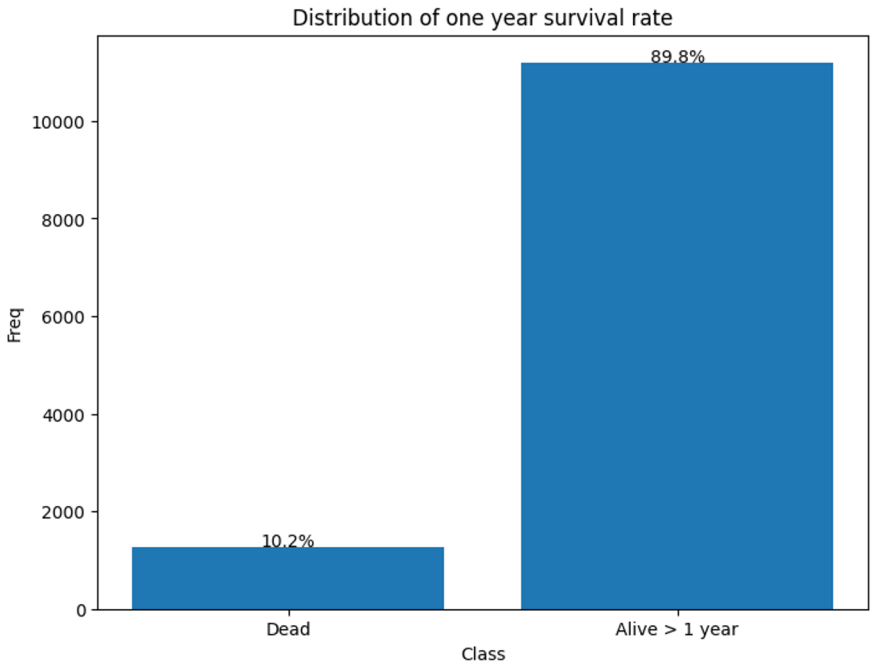
	wl_org	cod_wl	cod_ostxt_wl	num_prev_tx	thoracic_dgn	tah	vas	onvent	icu	inotropic	...	imuran_maint	cellcept_ind	cellcept_maint
0	HR	NaN	NaN	0	1201	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
1	HR	NaN	NaN	0	1000	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
2	HR	NaN	NaN	0	1007	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
3	HR	NaN	NaN	0	1049	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
4	HR	NaN	NaN	0	1999	N	N	N	N	NaN	...	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
120259	HR	NaN	NaN	0	1004	N	N	N	Y	NaN	...	1.0	NaN	NaN
120260	HR	NaN	NaN	0	1203	N	N	N	Y	N	...	0.0	NaN	NaN
120261	HR	NaN	NaN	0	1999	N	N	N	N	NaN	...	1.0	NaN	NaN
120262	HR	NaN	NaN	0	1000	NaN	NaN	NaN	NaN	NaN	...	NaN	0.0	1.0
120263	HR	NaN	NaN	0	1201	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN

Target Class Distribution

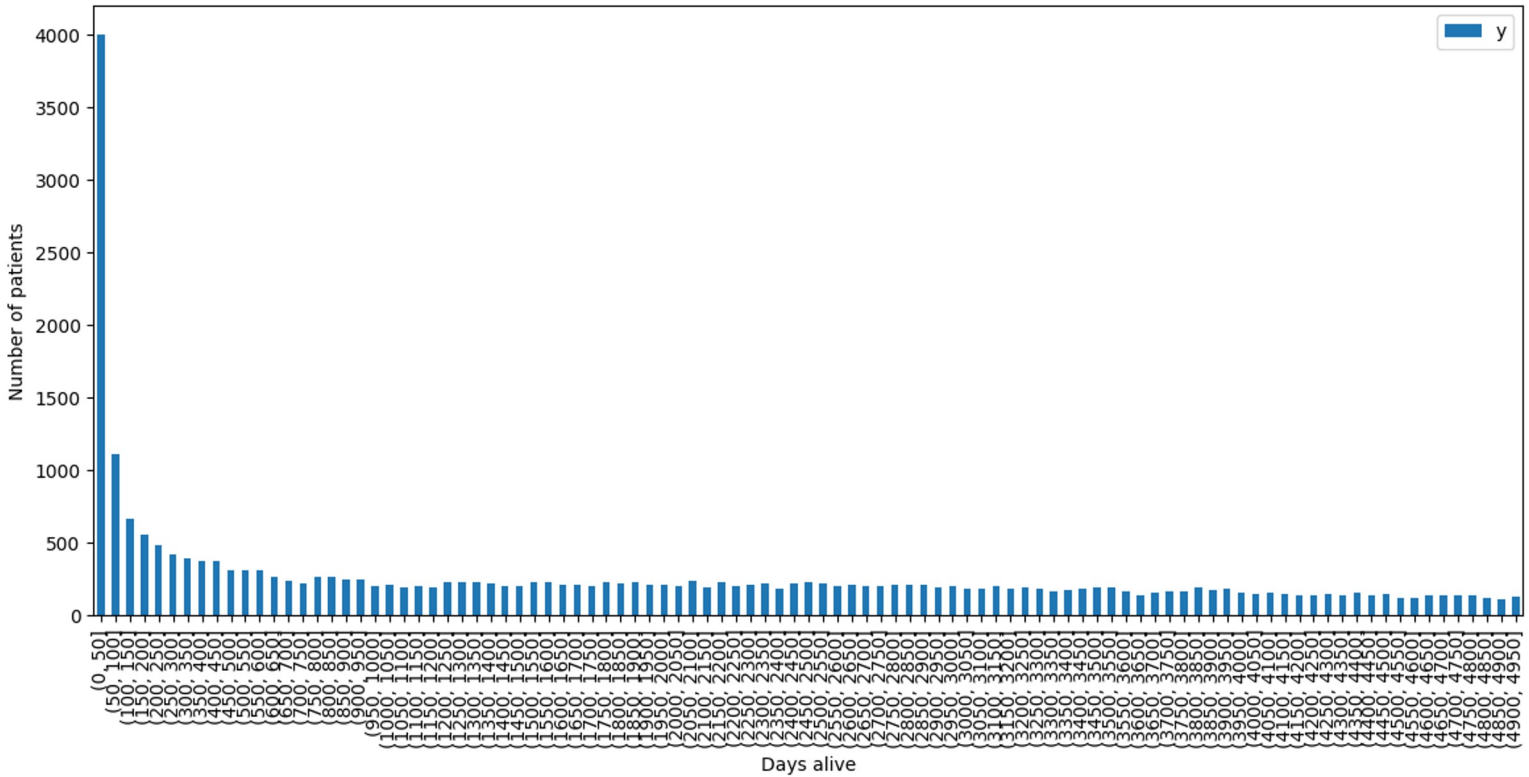
- Imbalanced dataset
  - Training data



  - Test data



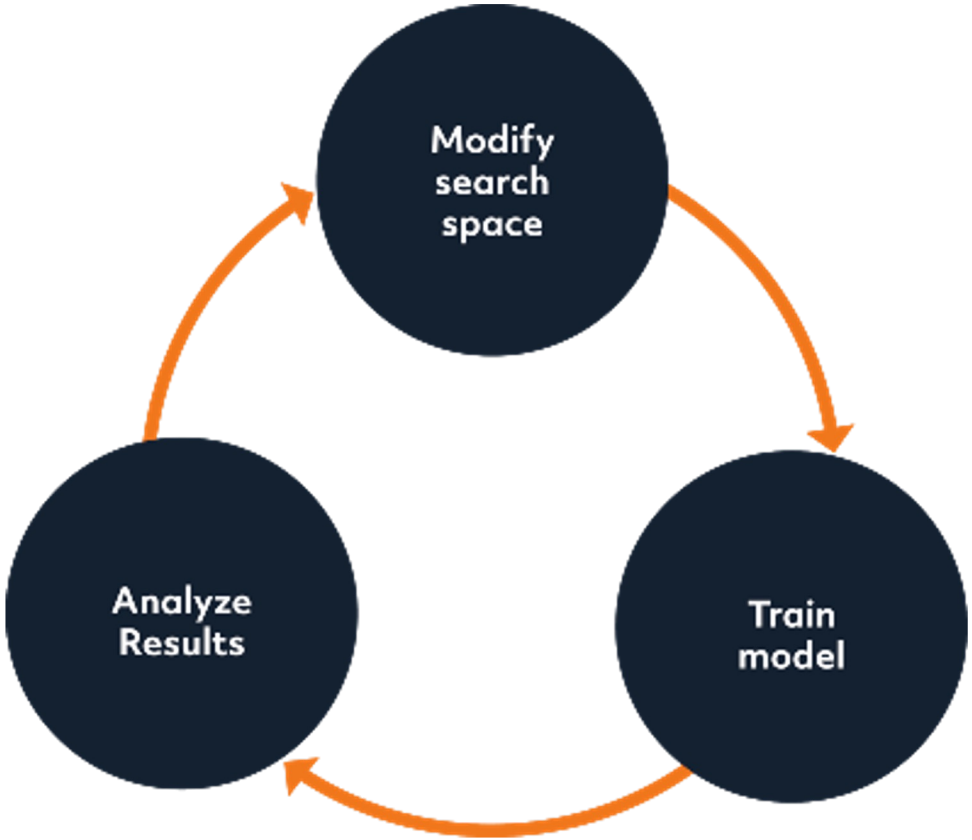
Time of death



Methodology

- Limitations
- Data handling

- Define model
- Train model
- Variable selection
- Model tuning
- Validating result

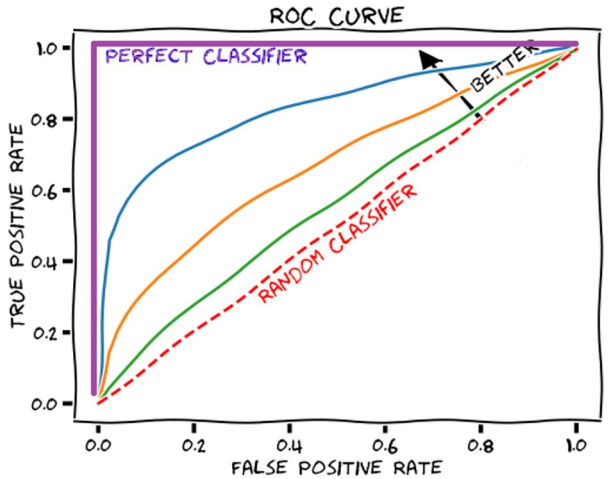


Evaluation metrics

- AUROC
- F1
- F1 MACRO

$$\text{F1 Score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

$$\text{Macro F1 Score} = \frac{\sum_{i=1}^{n} \text{F1 Score}_i}{n}$$



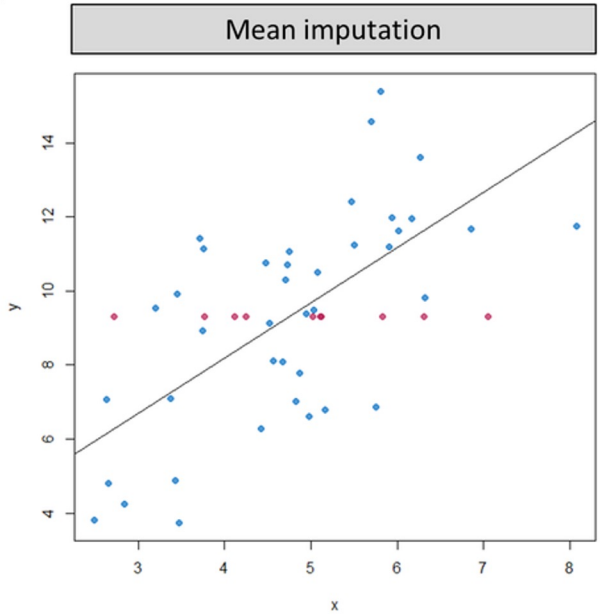
	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

Limitations

- Remove all patients under 18-years
- Remove columns with too many missing values
- Focused only on one-year survival

Data Handling

- Merge similar features
- Impute missing values
- Normalize data set
- One hot encoding categorical features
- Remove outliers
- Divide data into test and training data

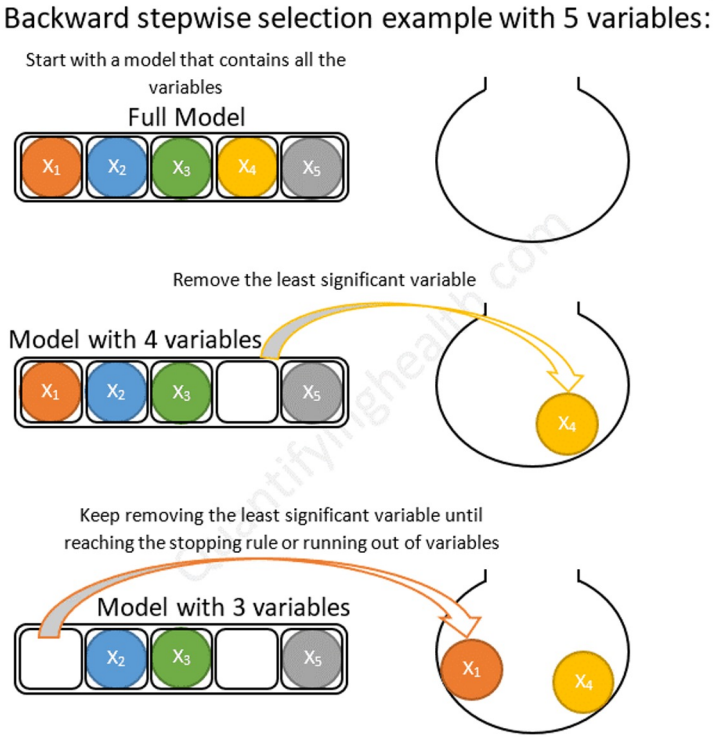


Data Handling - merging of columns

Cyklosporin_ind	Cyklosporin_maint
Cellcept_ind	Cellcept_maint
Tacrolimus_ind	Tacrolimus_maint
ATG_ind	ATG_maint
Kortison_ind	Kortison_maint
Imuran_ind	Imuran_maint
mTOR_ind	mTOR_maint
monoclonalAB_ind	monoclonalAB_maint
other_ind	other_maint
cig_use	

Variable Selection

- RFECV (Backward elimination)
- Cross validation
- Custom scorer (auc-roc)



Model Tuning

- Gridsearch

```
param_grid = {'C': [0.01, 0.1, 1.0],
              'penalty': ['l2'],
              'class_weight': [None, 'balanced']}
```

```
"C":0.1,
"penalty": "l2",
"class_weight": "balanced"
```

Model Result

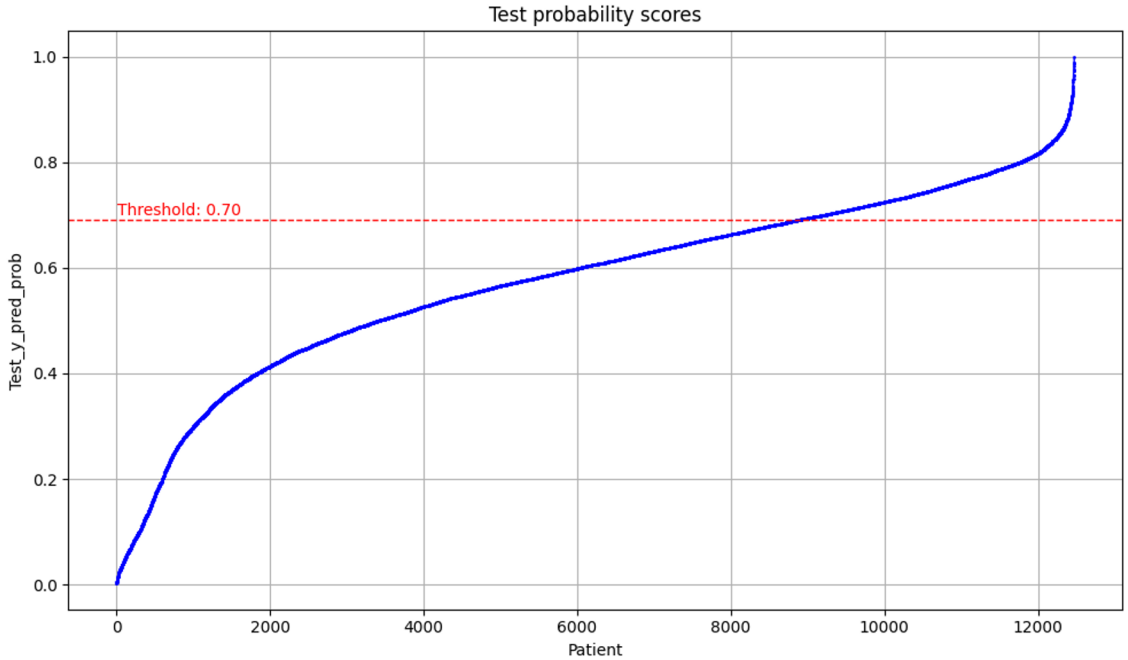
- **Logistic regression DX merged**

- **Neural Network - Keras - 64 nodes**
  - AUROC: 0.735

- **Random Forest**
  - AUROC: 0.72
  - F1 Macro: 0.59

Metric	Score
F1	0.947
F1_Macro	0.681
AUROC	0.740

Threshold


Test probability scores
Threshold: 0.70
Test_y_pred_prob
Patient
0.0
0.2
0.4
0.6
0.8
1.0
0
2000
4000
6000
8000
10000
12000

Best Model

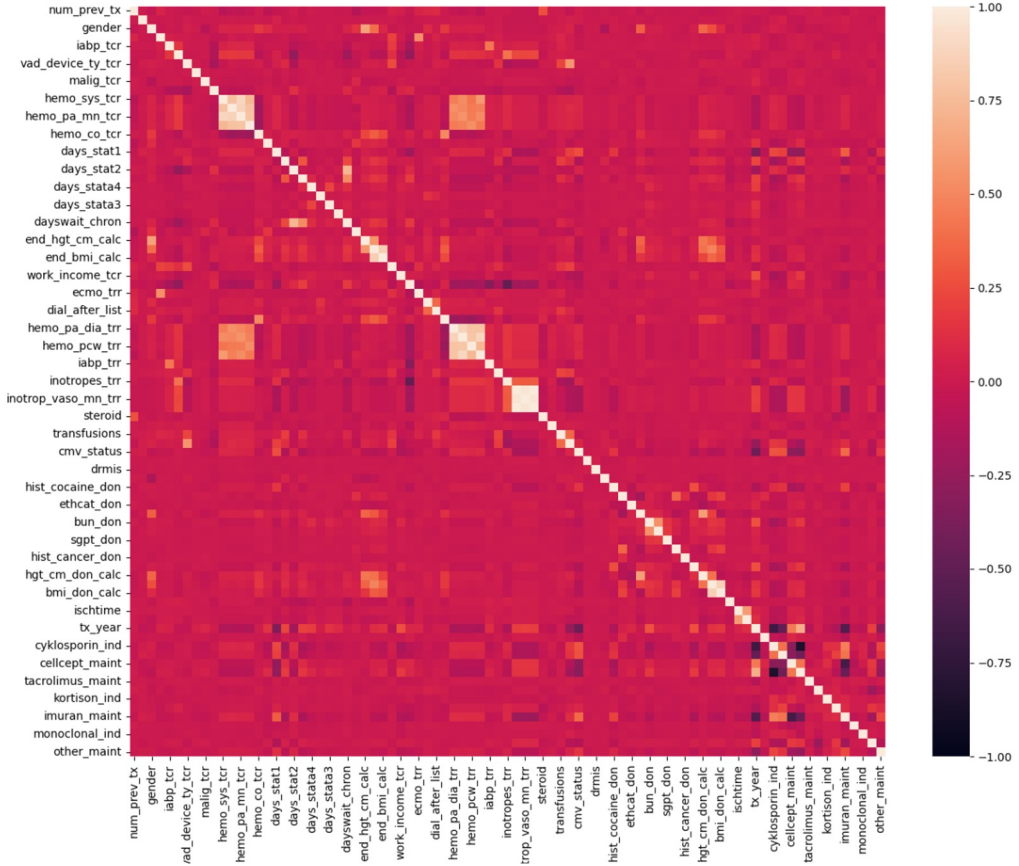
Parameters: {'C': 0.1, 'class_weight': 'balanced', 'dual': False, 'fit_intercept': True, 'intercept_scaling': 1, 'l1_ratio': None, 'max_iter': 100, 'multi_class': 'auto', 'n_jobs': None, 'penalty': 'l2', 'random_state': 42, 'solver': 'lbfgs', 'tol': 0.0001, 'verbose': 0, 'warm_start': False}

Explicit set parameters: {'random_state': 42, 'C': 0.1, 'penalty': 'l2', 'class_weight': 'balanced'}

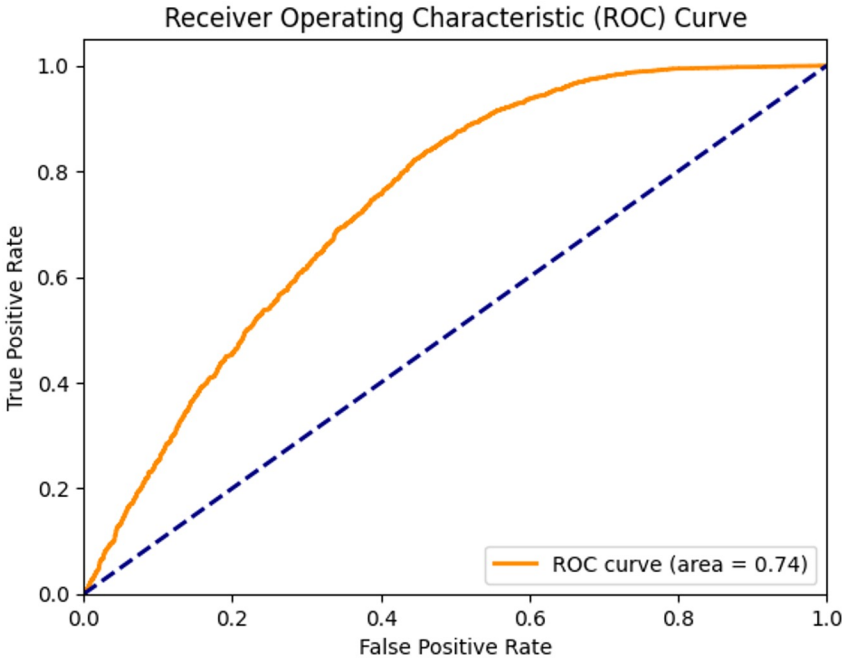
Number of Features: 89

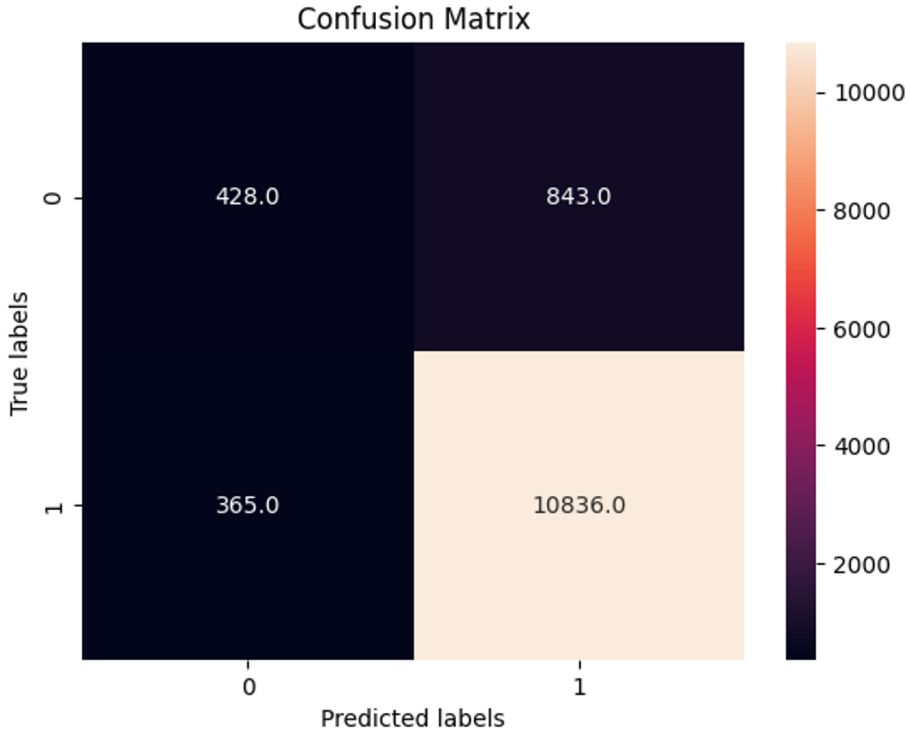
Features: ['num_prev_tx', 'gender', 'ecmo_tcr', 'iabp_tcr', 'inotropes_tcr', 'vad_device_ty_tcr', 'malig_tcr', 'most_rcnt_creat', 'tot_serum_album', 'impl_defibril', 'hemo_sys_tcr', 'hemo_pa_dia_tcr', 'hemo_pa_mn_tcr', 'hemo_pcw_tcr', 'hemo_co_tcr', 'prior_card_surg_tcr', 'days_stat1', 'days_stat1a', 'days_stat2', 'days_stat1b', 'days_stata4', 'days_stata5', 'days_stata2', 'days_stata3', 'days_stata1', 'days_stata6', 'dayswait_chron', 'init_age', 'end_hgt_cm_calc', 'end_wgt_kg_calc', 'end_bmi_calc', 'ventilator_tcr', 'work_income_tcr', 'med_cond_trr', 'ecmo_trr', 'creat_trr', 'hemo_pa_dia_trr', 'hemo_pa_mn_trr', 'hemo_pcw_trr', 'hemo_sys_trr', 'iabp_trr', 'infect_iv_drug_trr', 'inotropes_trr', 'inotrop_vaso_co_trr', 'inotrop_vaso_dia_trr', 'inotrop_vaso_pcw_trr', 'steroid', 'tbili', 'transfusions', 'vad_device_ty_trr', 'cmv_status', 'amis', 'bmis', 'drmis', 'hlamis', 'pramr_cl1', 'hist_cocaine_don', 'age_don', 'ethcat_don', 'gender_don', 'bun_don', 'creat_don', 'sgot_don', 'sgpt_don', 'tbili_don', 'hist_hypertens_don', 'hist_cancer_don', 'hist_oth_drug_don', 'hgt_cm_don_calc', 'wgt_kg_don_calc', 'bmi_don_calc', 'abo_mat', 'ischtime', 'distance', 'tx_year', 'pulm_cath_don', 'cyklosporin_ind', 'cyklosporin_maint', 'cellcept_maint', 'tacrolimus_ind', 'tacrolimus_maint', 'atg_maint', 'kortison_ind', 'kortison_maint', 'imuran_maint', 'mtor_ind', 'monoclonal_ind', 'monoclonal_maint', 'other_maint']

Logistic Regression

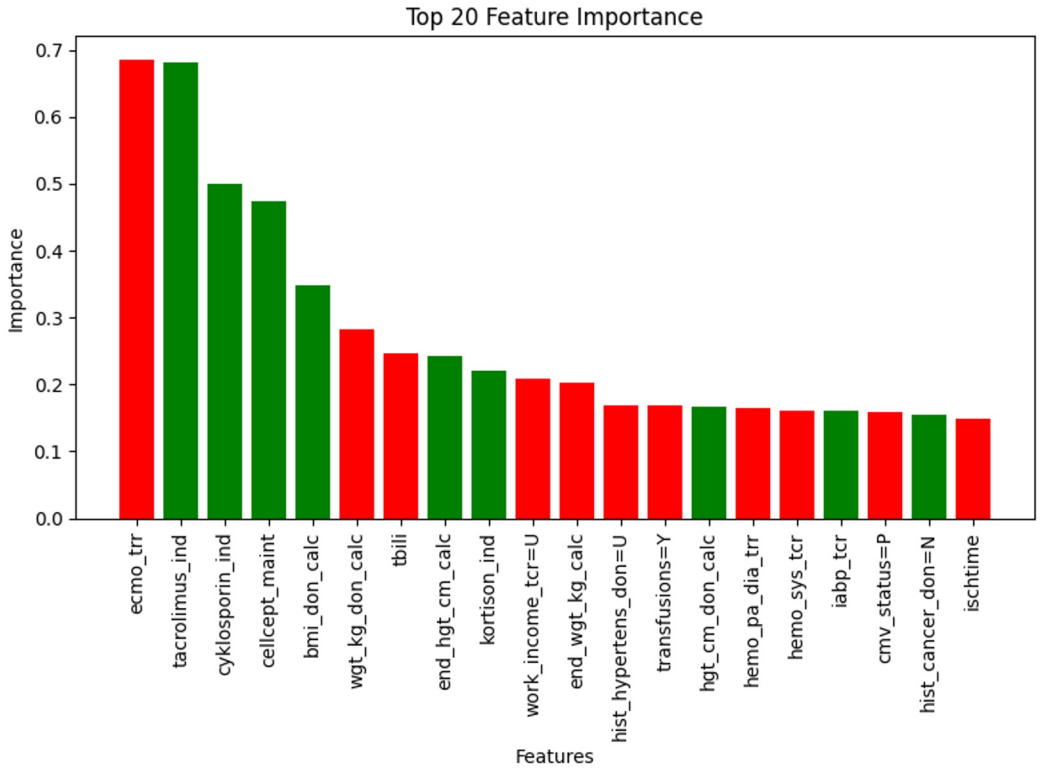


Logistic Regression





Logistic Regression



Logistic Regression

- Lower score - important feature
  - Tacrolimus.. etc

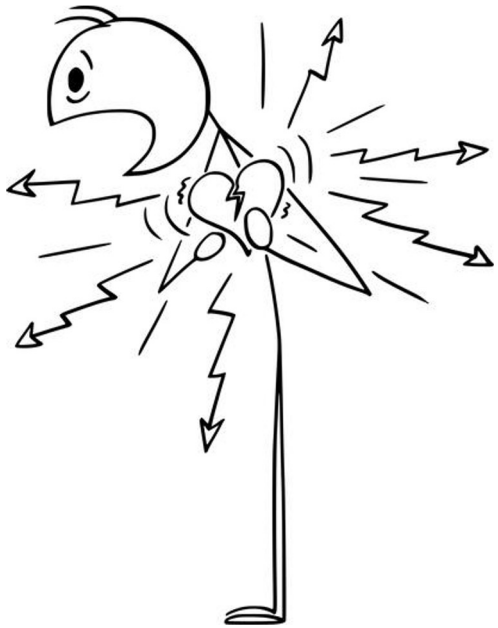
Rank	Feature
tacrolimus_ind	0.7213270058289025
cyklosporin_ind	0.7306819927494672
tbili	0.7343246791989391
ischtime	0.7356312529980217
age_don	0.737007928439569
tot_serum_album	0.737974108892576
creat_trr	0.7381942477177104
cellcept_maint	0.7382828932816285
kortison_ind	0.738561052103432
init_age	0.7385966648616782
end_hgt_cm_calc	0.7388083746316064
distance	0.7390164318109452
vad_device_ty_trr	0.7392306000553088
days_stata4	0.7392851781877686
tacrolimus_maint	0.7394182870178994
num_prev_tx	0.7394372523921132
hlamis	0.7394564284927073
days_stata5	0.7394601513254233
monoclonal_maint	0.7394817859004524
pramr_cl1	0.7394823478374661

Conclusion

- Goal, most important features:
  - Medications +
  - No history of cancer +
  - Ecmo -

- Best performing model
  - Logistic regression
  - Auc-roc score: 0.740

- IHTSA model (2009–2011)
  - Auc-roc score: 0.654



LUND
UNIVERSITY