

KAGGLE CASE

CARDIOVASCULAR DISEASE DATASET



<https://github.com/Victorbs18/KAGGLE>

VÍCTOR BENITO SEGURA

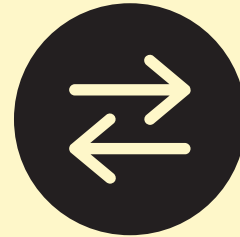
ORIGINAL FEATURES

- id
- gender
- height
- weight
- ap_hi
- ap_lo
- cholesterol
- gluc
- smoke
- alco
- active
- cardio

Shape: 70.000 rows, 12 columns

FEATURE ENGINEERING:

TRANSFORMATIONS:



- **gender:**
- **cholesterol:** normal, above, well_above
- **glucosa:** normal, above, well_above
- **age:** days -> years, young_adult, adult, old

DELETED FEATURES:

- **id**



NEW FEATURES:



- **BMI:** underweight, healthy, overweight, obese, severely_obese, morbidly_obese

23 FEATURES 1 TARGET VALUE

DATA CLEANING:



NULL-VALUES: no null values

OUTLIERS

- **height:** log, 99% quantile
- **weight:** log, 99% quantile
- **ap_hi:** >0, <250
- **ap_lo:** >0, <200

OUTLIERS REMOVED

0
349
1188
3494

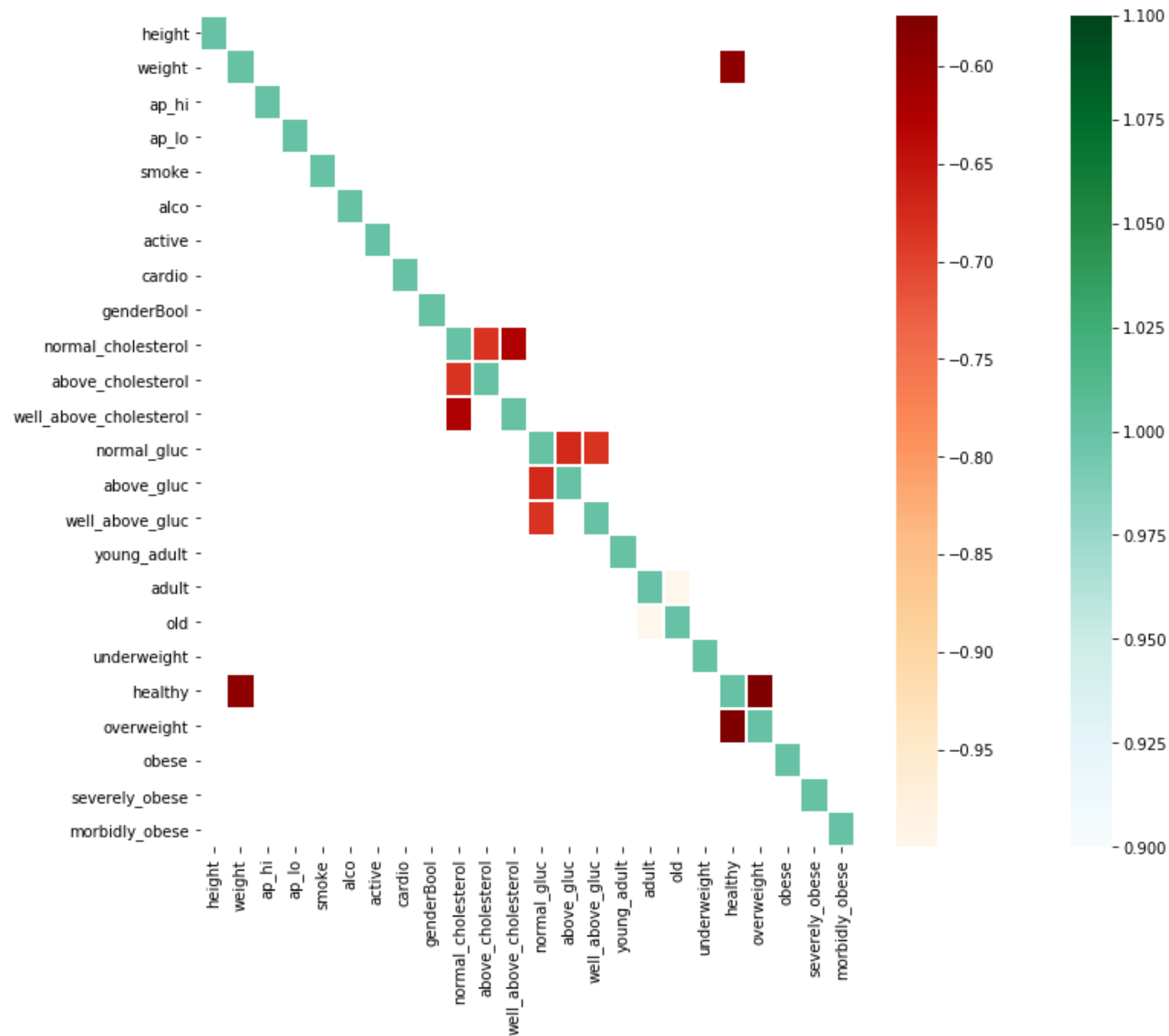
67394 rows, 3.7% rows removed

TARGET FEATURE

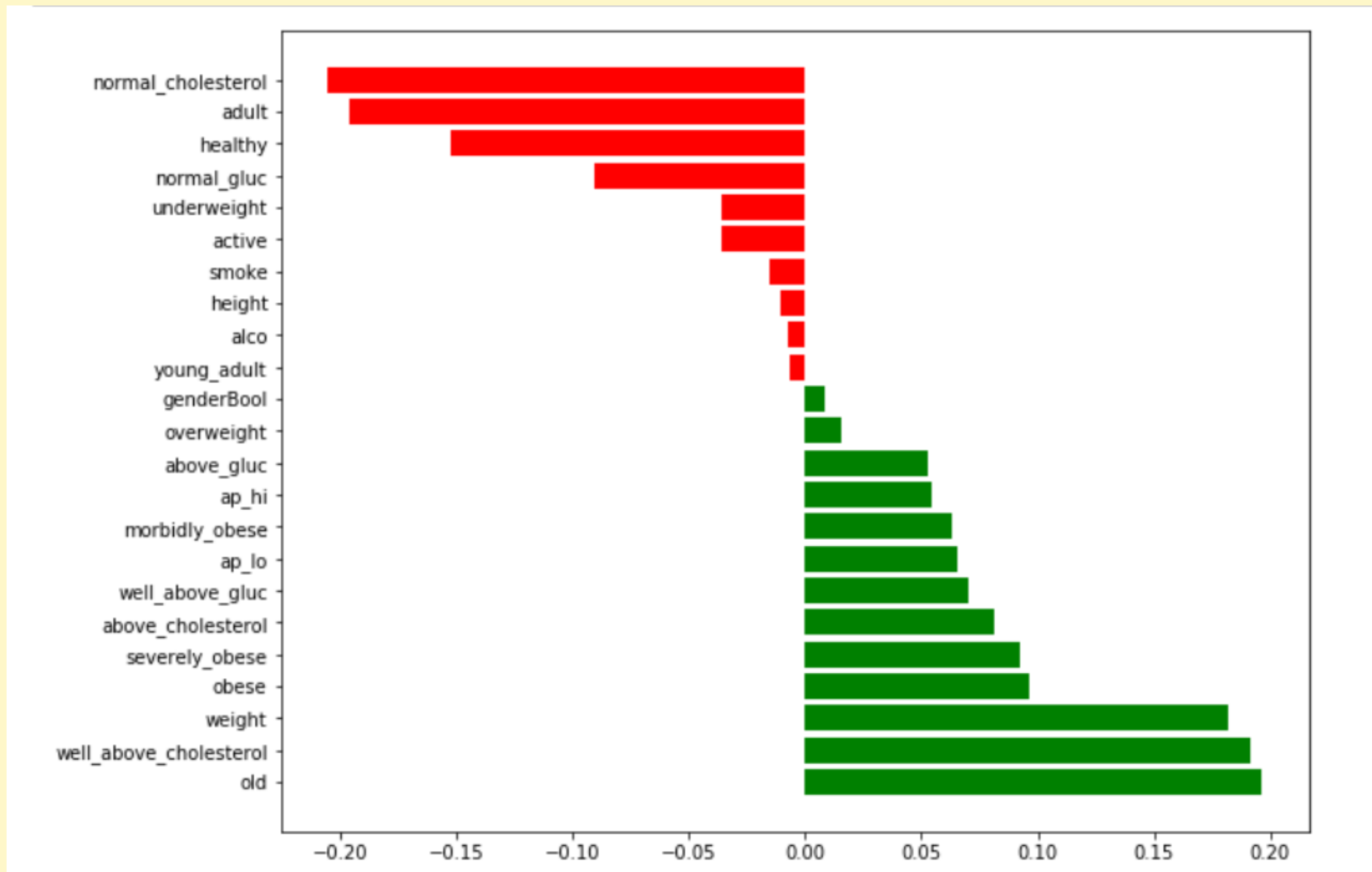


cardio

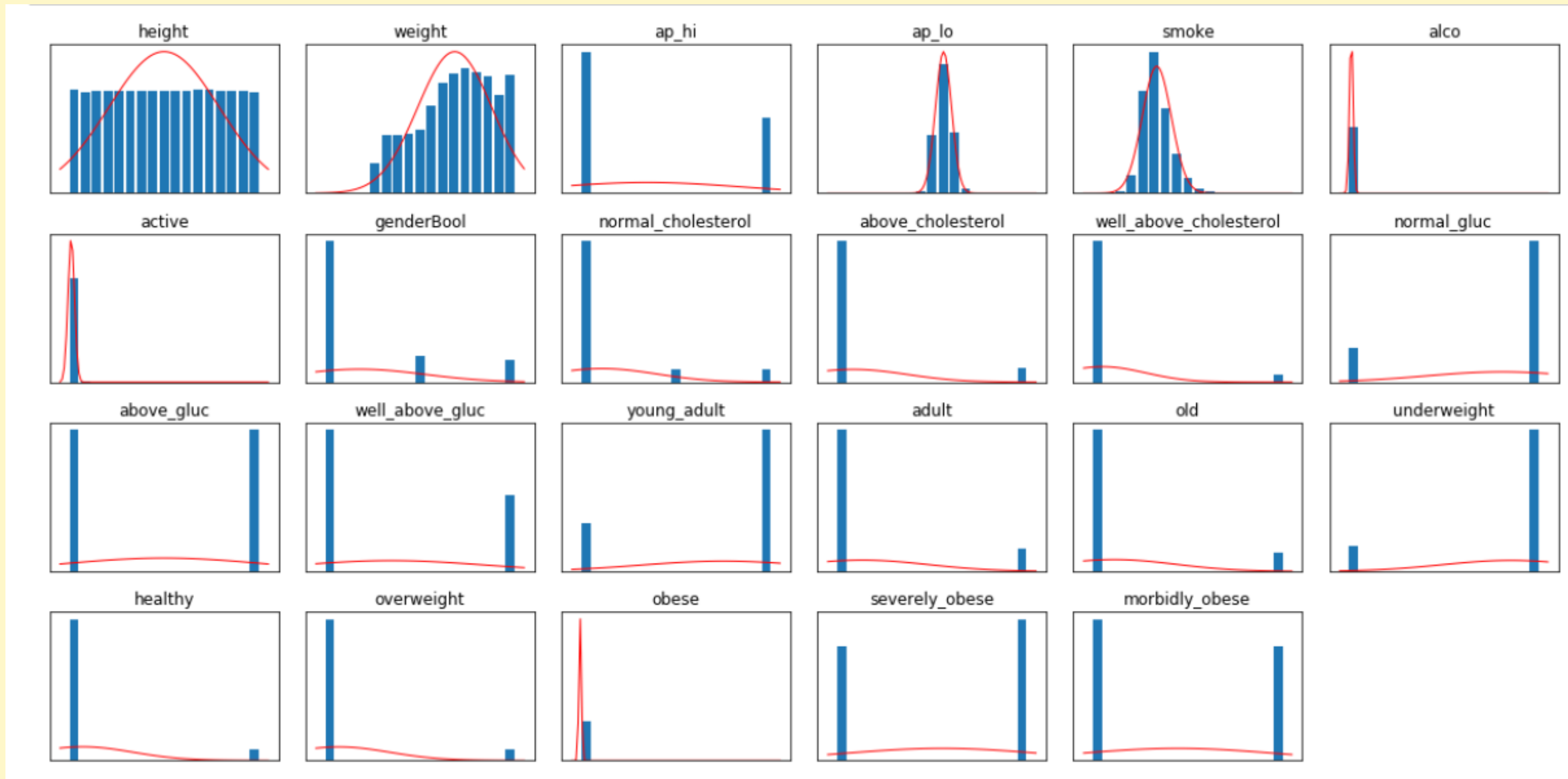
CORRELATION MATRIX



CORRELATION WITH TARGET

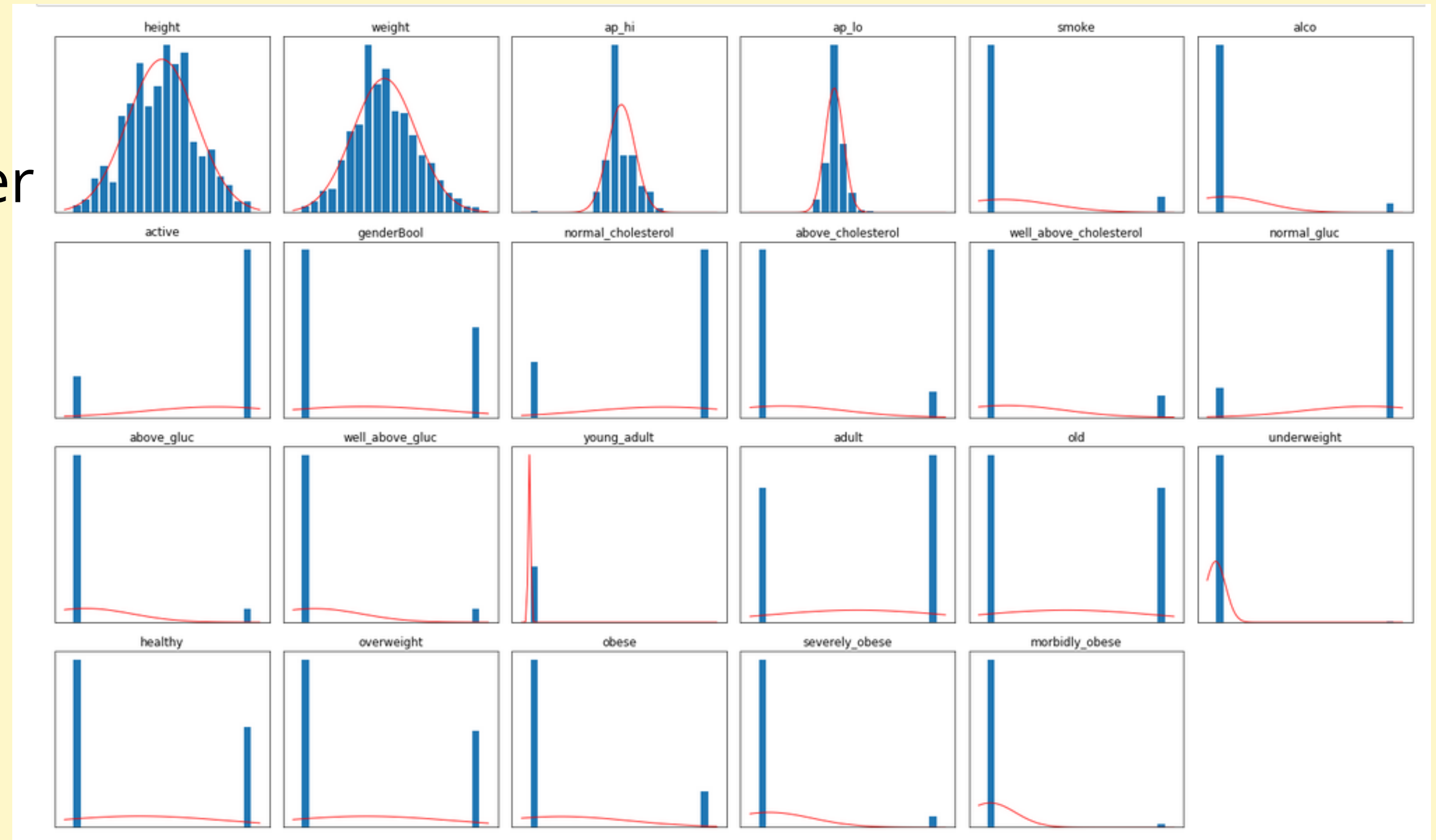


FEATURES DISTRIBUTION



NORMALIZED FEATURES DISTRIBUTION

- StandardScaler
- Normalizer



BALANCED DATASET



49.5% of rows with 'cardio'=1

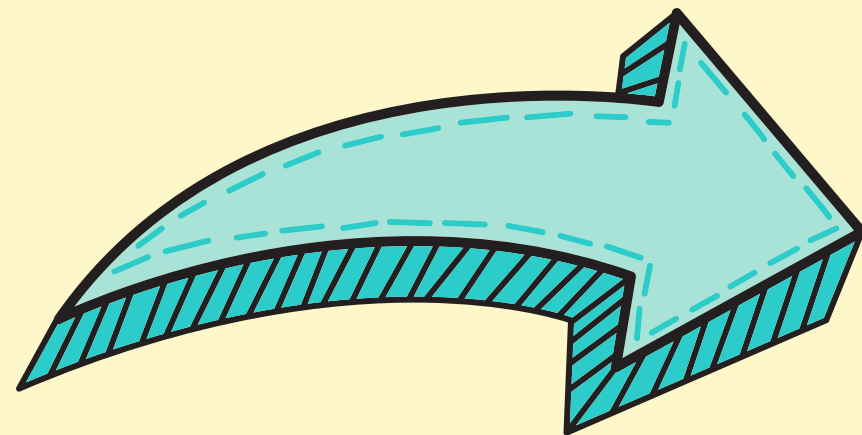
FEATURE SELECTION

RECURSIVE FEATURE ELIMINATION (RFE): ap_hi, well_above_cholesterol, old, 0.7297.

LASSO: height, weight, ap_hi, severely_obese, morbidly_obese, 0.711.

POLYNOMIAL FEATURES: 23 -> 299 features.

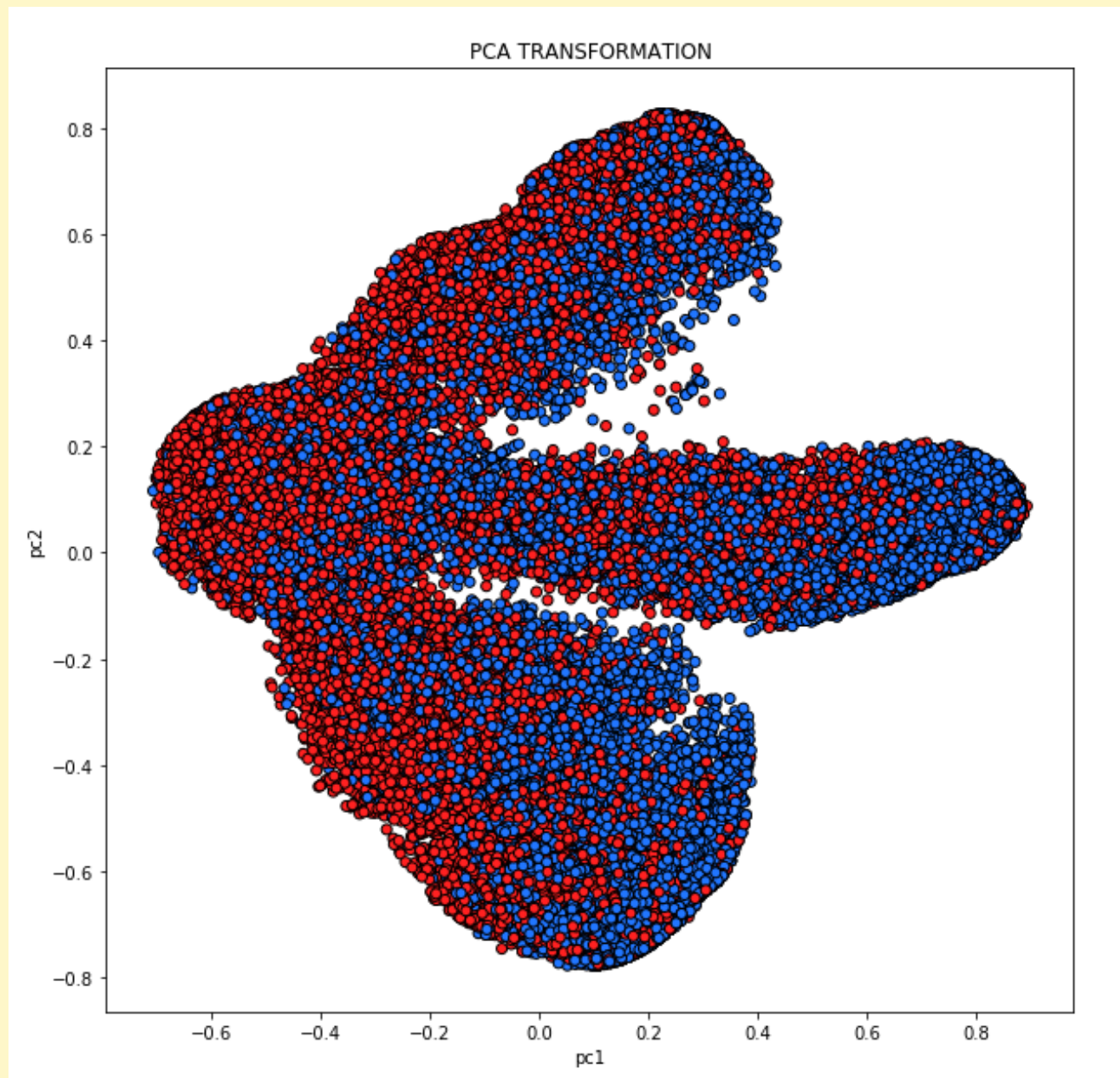
SELECTKBEST: ap_hi, ap_lo, normal_cholesterol, well_above_cholesterol, adult, 0.725



RFE

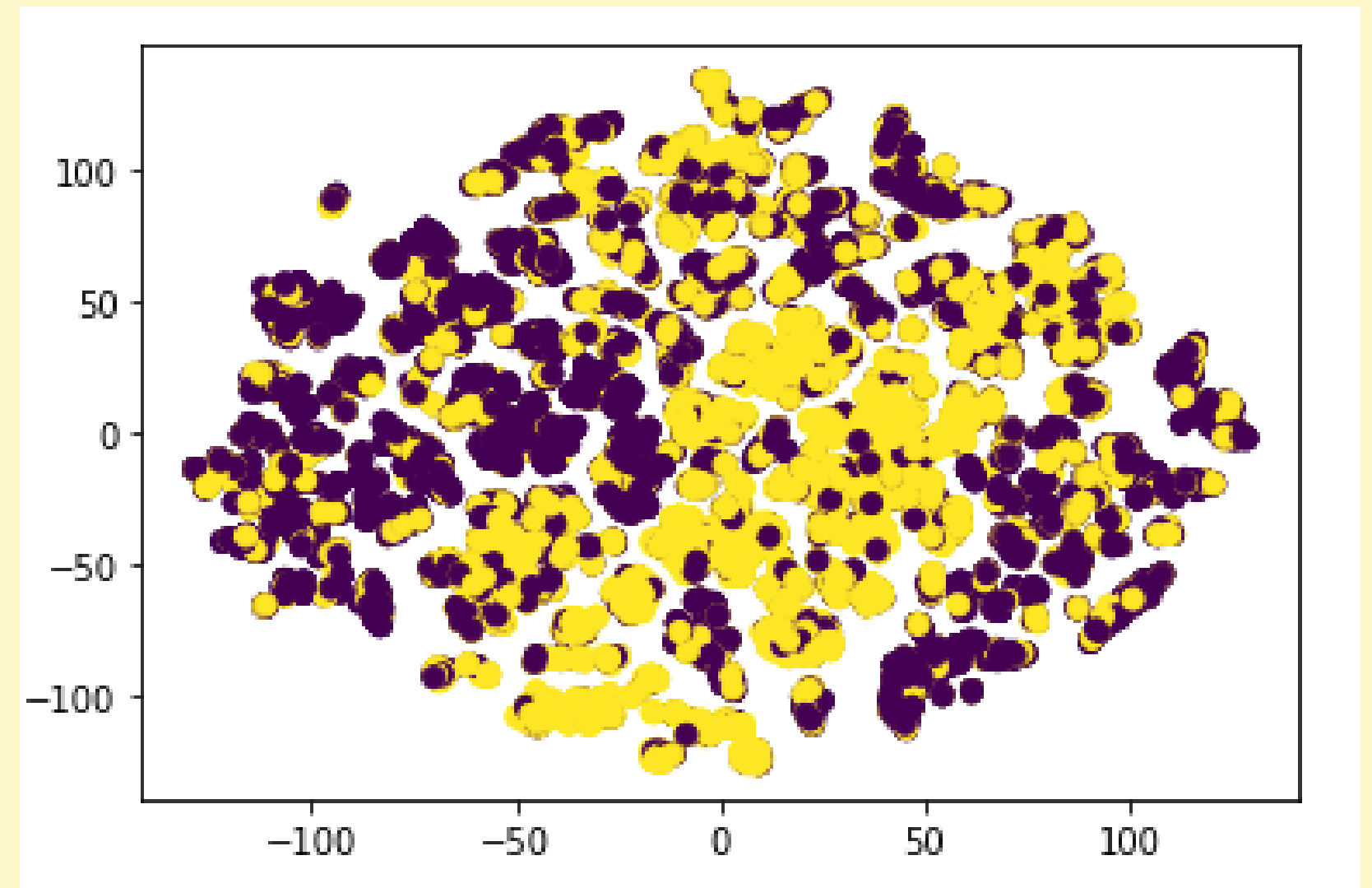
PCA

2 components, 0.662



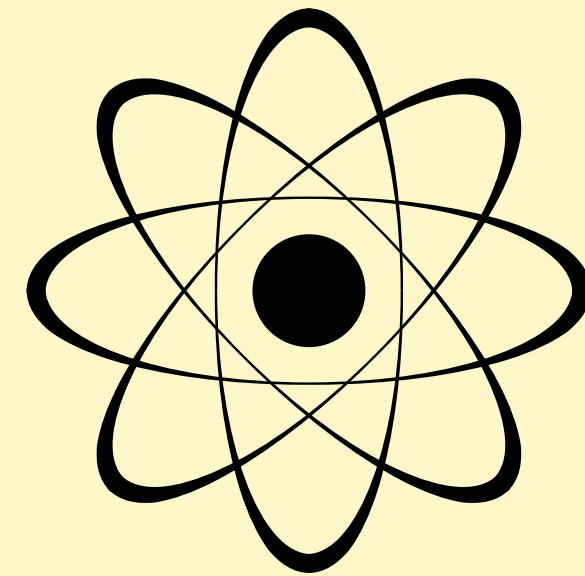
TSNE

2 components, reduced: 20k
0.589

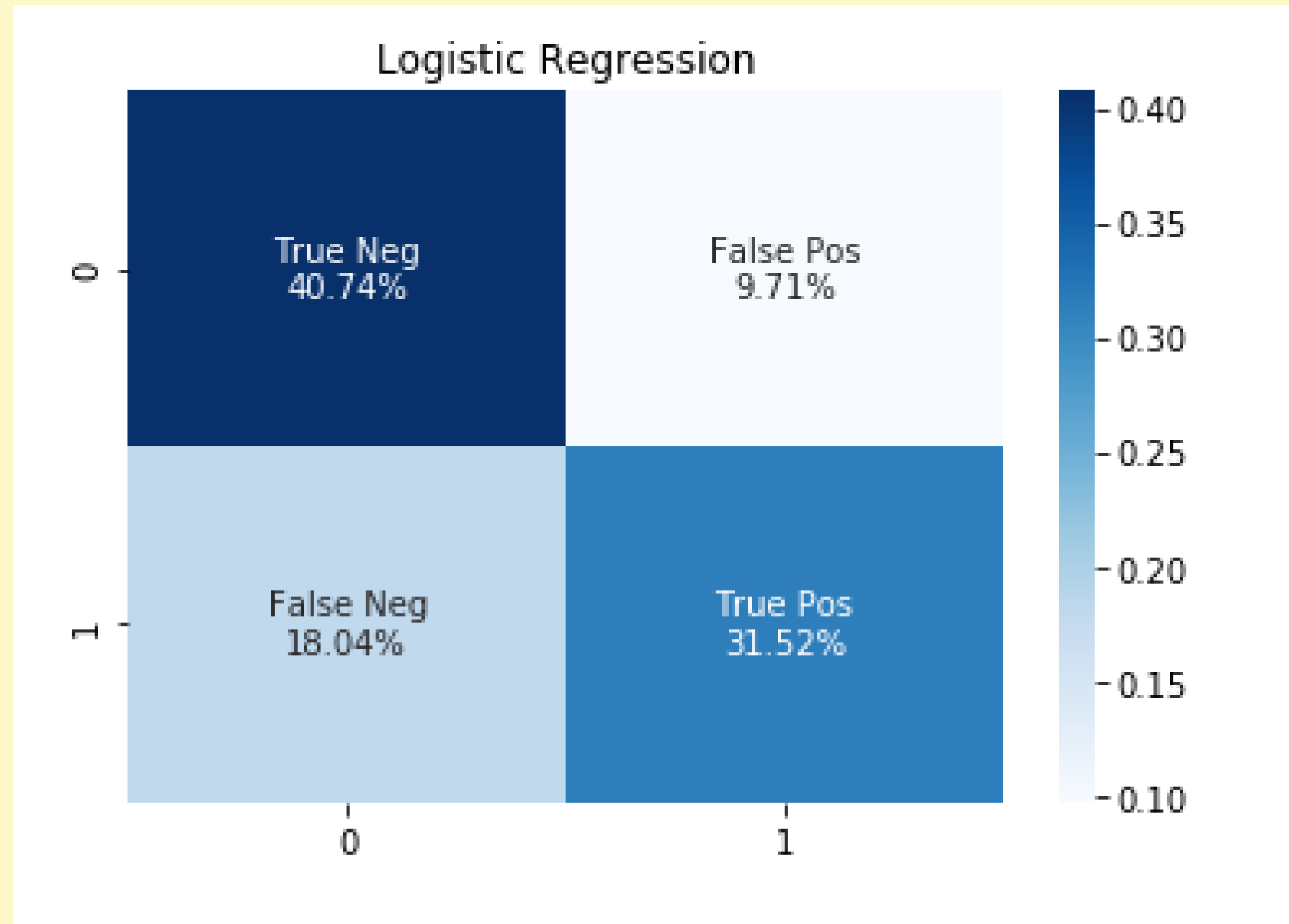


MODELS

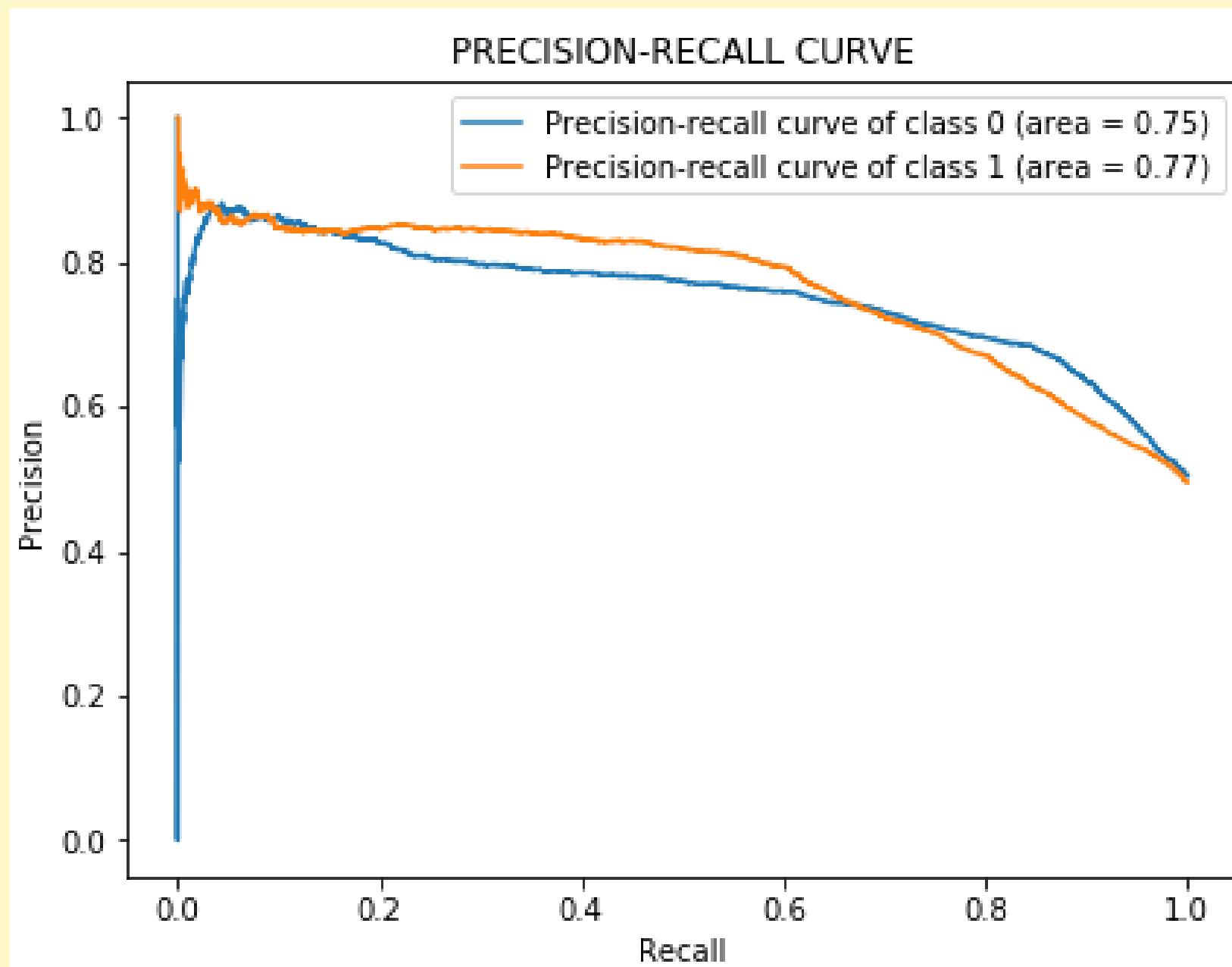
- 1 - LOGISTIC REGRESSION
- 2- LINEAR SVC
- 3- KNN
- 4- NAIVE BAYES
- 5- RANDOM FOREST
- 6- DECISION TREES
- 7- SGDCCLASSIFIER
- 8- LINEAR DISCRIMINANT ANALYSIS
- 9- XGBOOST
- 10-PERCEPTRON



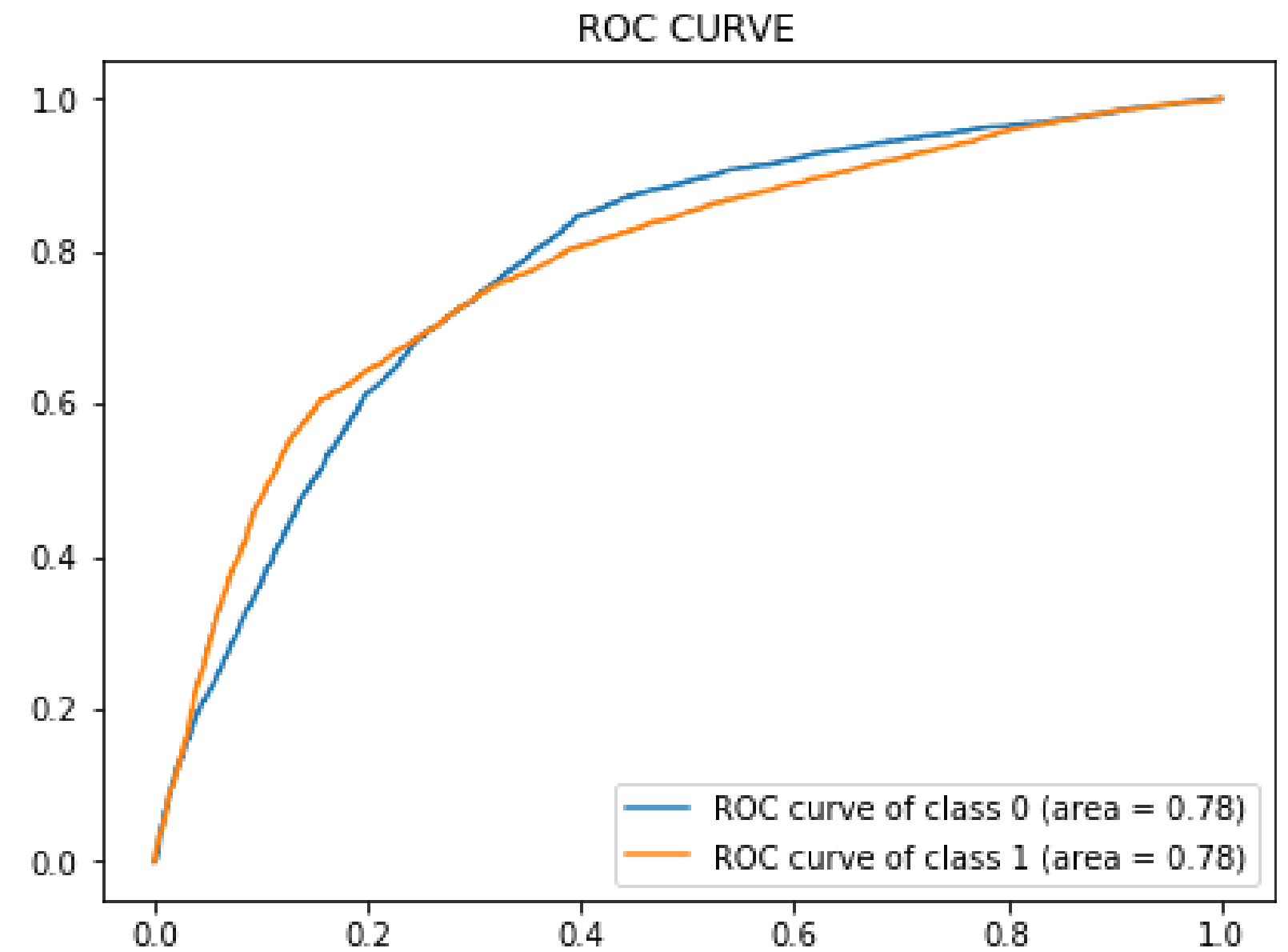
CONFUSION MATRIX



PRECISION-RECALL CURVE

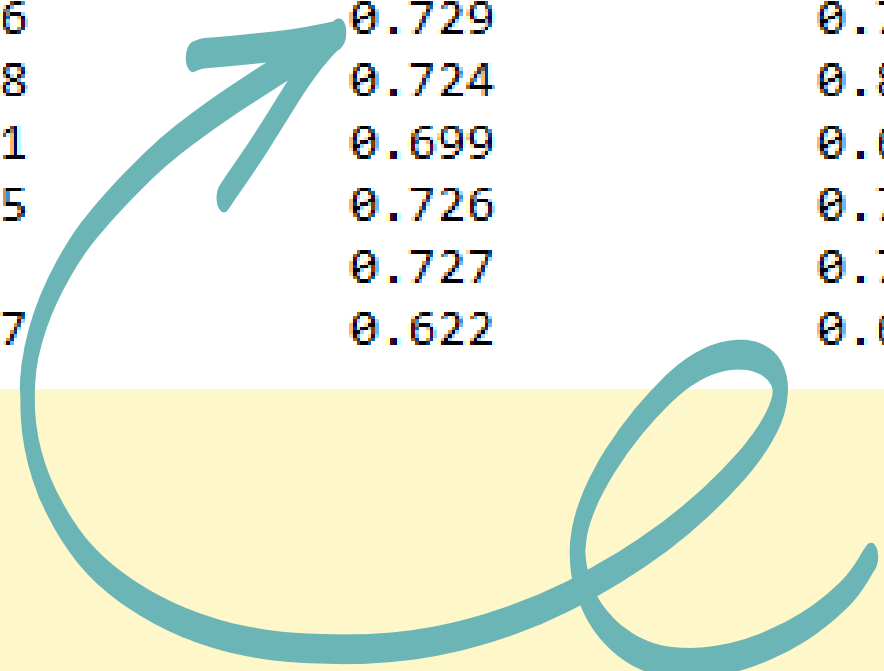


ROC CURVE



RESUME MODELS ALL FEATURES

Model	Train Accuracy	Test Accuracy	Recall	Precision	F1	Average precision	Elapsed Time
Logistic Regression	0.726	0.728	0.767 0.689	0.713 0.745	0.739 0.716	0.668	1.575
Linear SVC	0.725	0.726	0.768 0.684	0.71 0.745	0.738 0.713	0.667	32.36
KNN	0.7	0.689	0.71 0.667	0.683 0.696	0.696 0.681	0.63	17.01
Naive Bayes	0.679	0.663	0.514 0.813	0.735 0.624	0.605 0.706	0.6	0.2493
Random Forest	0.726	0.729	0.788 0.669	0.706 0.758	0.745 0.71	0.672	42.02
Decision Tree	0.728	0.724	0.822 0.624	0.688 0.777	0.749 0.692	0.672	3.669
SGDClassifier	0.701	0.699	0.69 0.708	0.705 0.694	0.697 0.701	0.637	0.572
Linear Discriminant	0.725	0.726	0.768 0.683	0.71 0.745	0.738 0.713	0.667	1.058
XGBoost	0.73	0.727	0.785 0.669	0.705 0.755	0.743 0.71	0.67	35.08
Perceptron	0.637	0.622	0.67 0.573	0.613 0.633	0.64 0.601	0.575	0.3473



RANDOM FOREST: 0.729

RESUME MODELS WITH FEATURE SELECTION(RFE)

Model	Train Accuracy	Test Accuracy	Recall	Precision	F1	Average precision	Elapsed Time			
Logistic Regression	0.722	0.723	0.807	0.636	0.693	0.764	0.746	0.694	0.667	0.4561
Linear SVC	0.722	0.724	0.825	0.62	0.689	0.777	0.751	0.69	0.67	21.2
KNN	0.7	0.699	0.733	0.665	0.69	0.71	0.711	0.686	0.638	2.629
Naive Bayes	0.712	0.715	0.863	0.565	0.669	0.802	0.754	0.663	0.669	0.177
Random Forest	0.724	0.726	0.809	0.641	0.696	0.767	0.748	0.698	0.67	18.02
Decision Tree	0.725	0.725	0.826	0.622	0.69	0.779	0.752	0.692	0.672	0.6067
SGDClassifier	0.684	0.702	0.706	0.699	0.705	0.7	0.705	0.699	0.638	0.4458
Linear Discriminant	0.722	0.723	0.826	0.619	0.688	0.777	0.75	0.689	0.67	0.2824
XGBoost	0.724	0.727	0.803	0.65	0.7	0.764	0.748	0.703	0.67	8.877
Perceptron	0.587	0.637	0.709	0.564	0.623	0.656	0.663	0.606	0.586	0.2554



XGBOOST: 0.727

CONCLUSIONS

- SELECTED MODELS
- REASONS FOR A LOW ACCURACY
- FUTURE WORK

