Wine Quality

Predictive Model

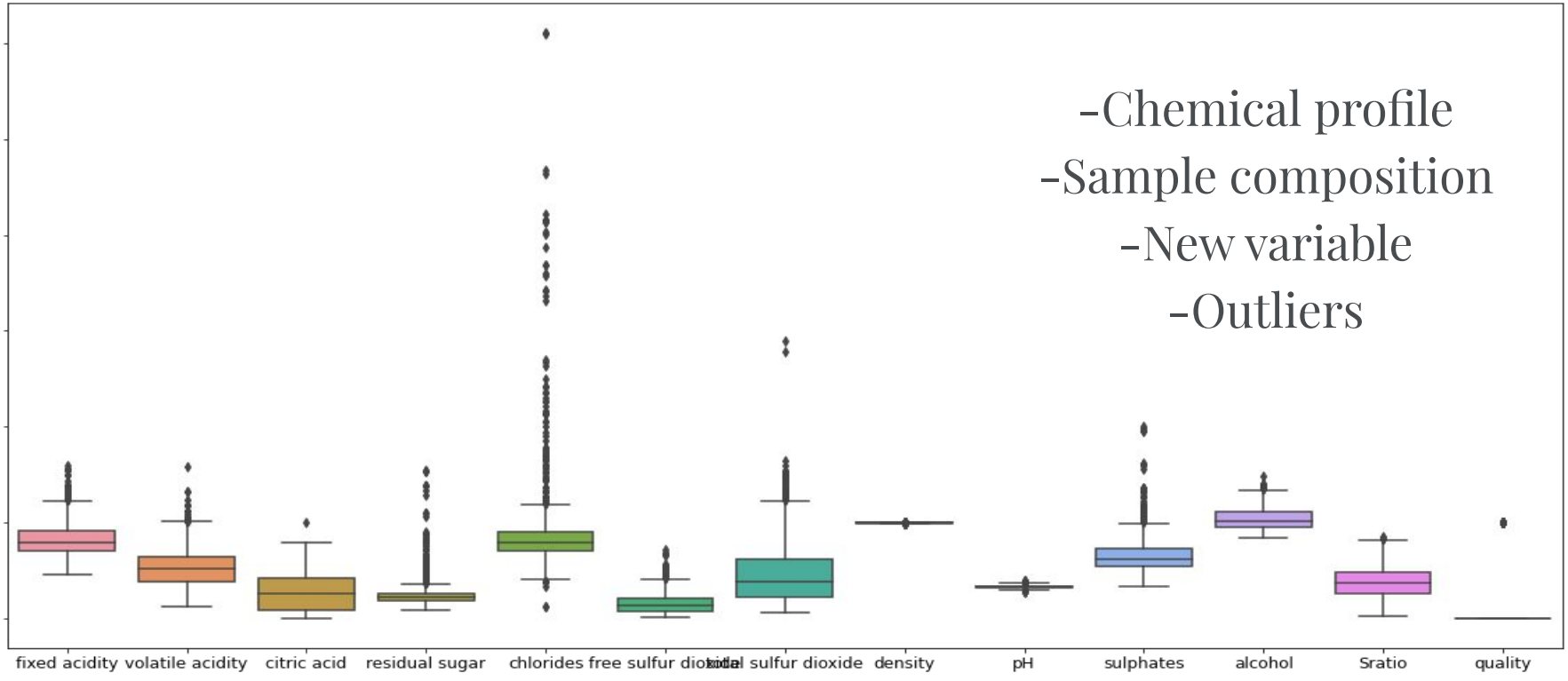
Christos Logaras
Dimitris Papadopoulos

Big Blue
DATA ACADEMY

Confusion Matrix

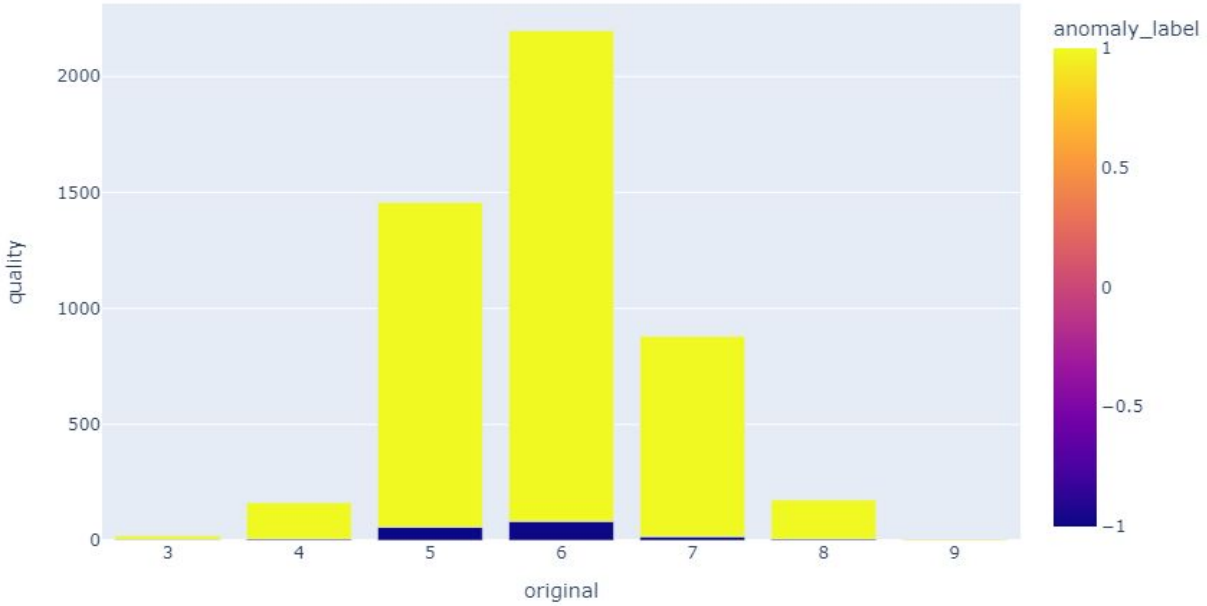
	$\hat{Y}=0$ Negative	$\hat{Y}=1$ Positive
$Y=0$ Inferior		
$Y=1$ Superior	on sale 	

EDA

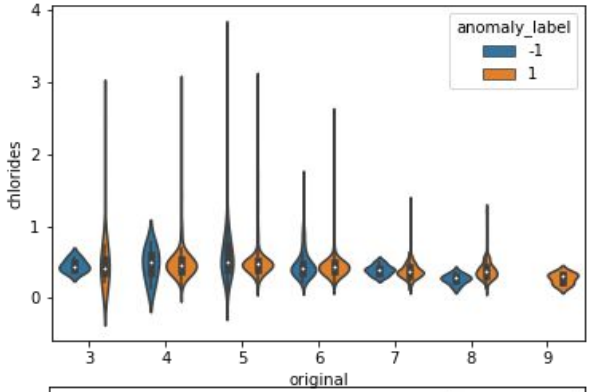


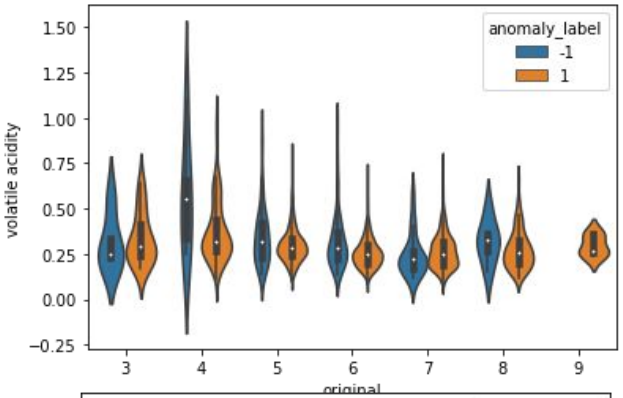
-Chemical profile
-Sample composition
-New variable
-Outliers

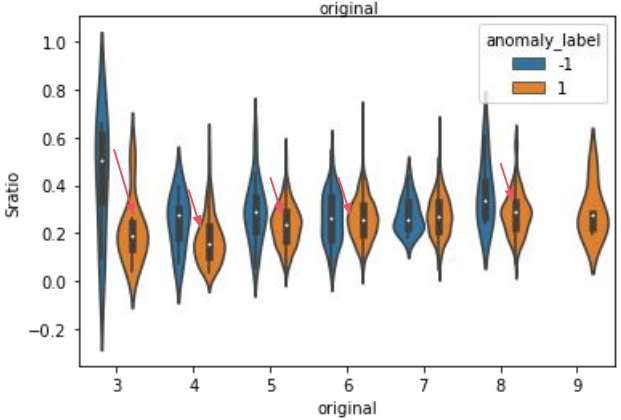
Outliers

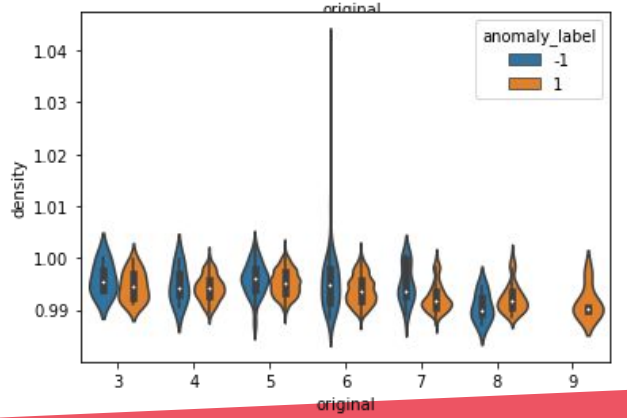

anomaly_label
1
0.5
0
−0.5
−1
quality
2000
1500
1000
500
0
3
4
5
6
7
8
9
original

Outliers

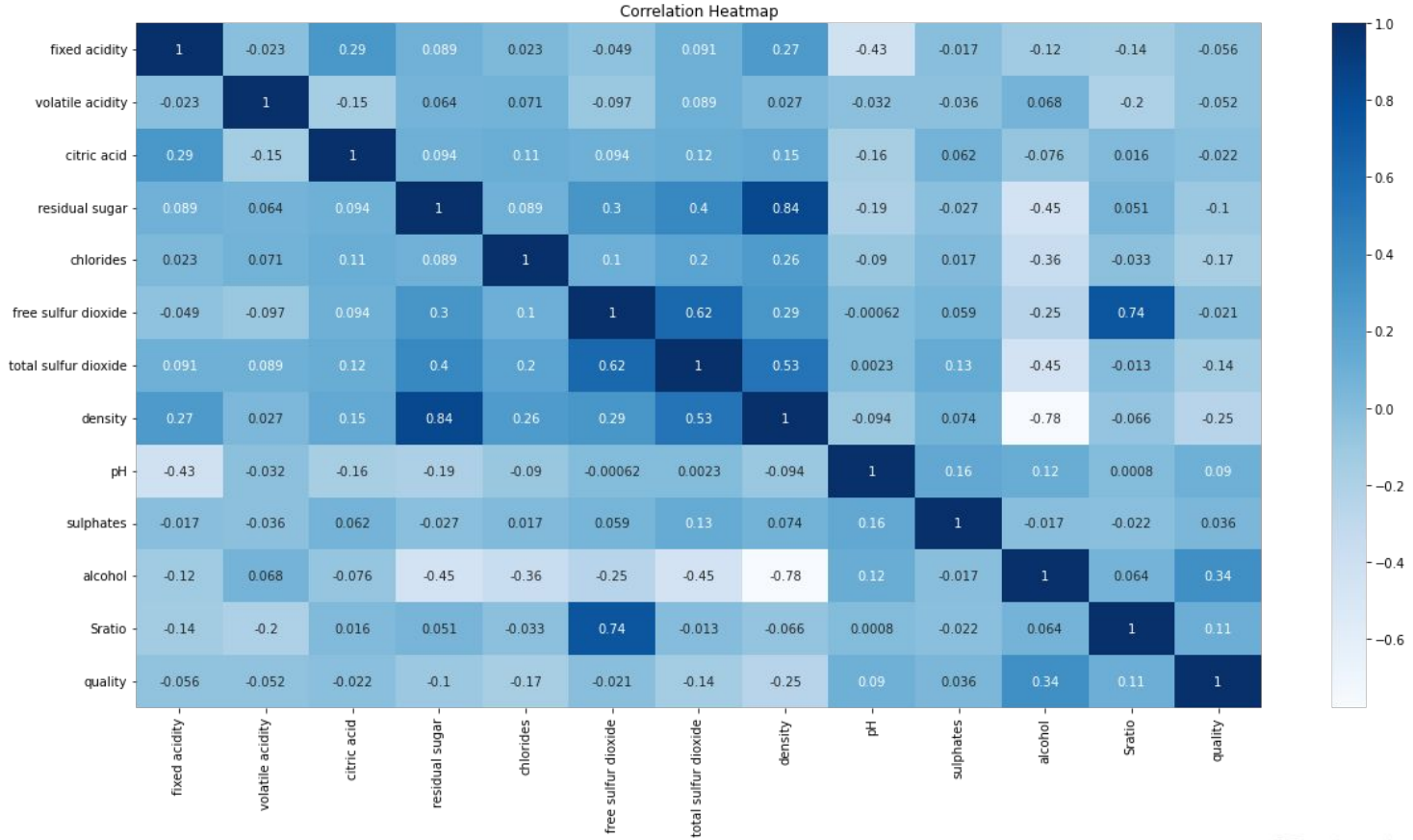

anomaly_label
-1
1
chlorides
original


anomaly_label
-1
1
volatile acidity
original

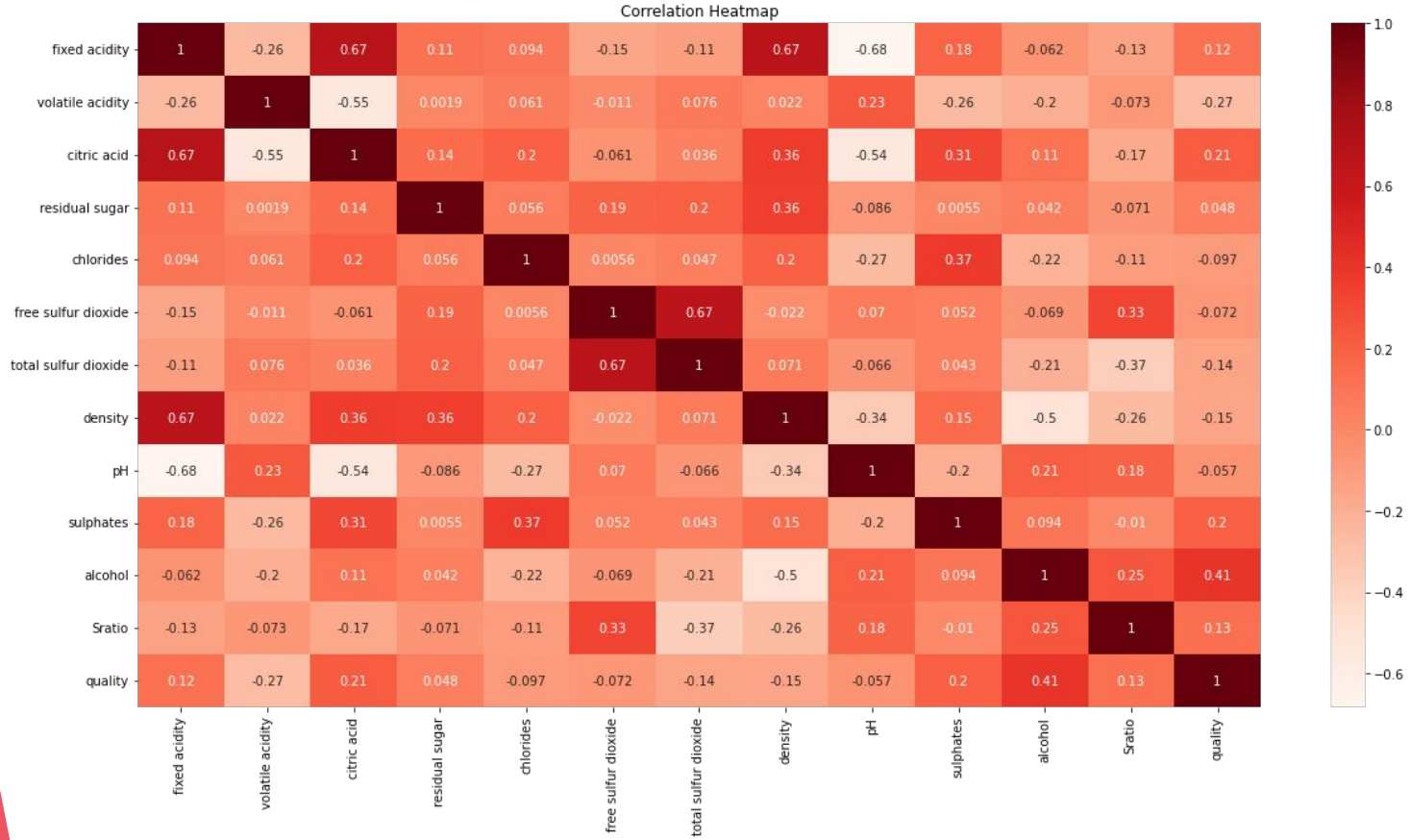

anomaly_label
-1
1
Sratio
original


anomaly_label
-1
1
density
original

Correlation Heatmaps



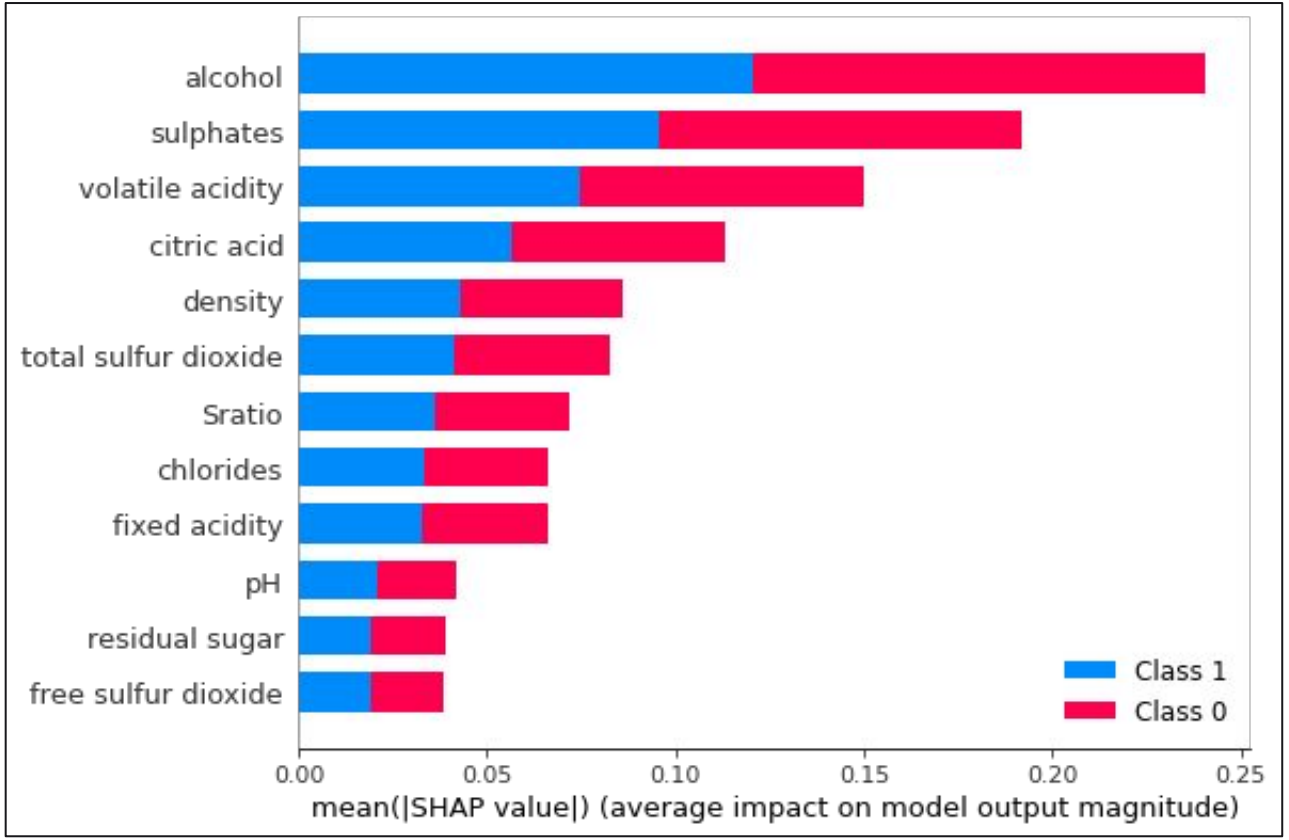
Correlation Heatmaps



Scallers

Classifier	Scaller	Stratify	Score	F1 1	Precision 1	Recal 1	test average precision	test balanced accuracy	test roc auc
LogisticRegression		√	0,8	0,31	0,59	0,21	0,54	0,59	0,79
LogisticRegression	StandardScaler	√	0,8	0,38	0,59	0,28	0,54	0,61	0,8
LogisticRegression	MinMaxScaler	√	0,8	0,33	0,58	0,23	0,54	0,6	0,79
LogisticRegression	RobustScaler	√	0,8	0,36	0,58	0,26	0,54	0,61	0,8
LogisticRegression	Normalizer	√	0,79	0,15	0,58	0,08	0,52	0,54	0,78
LogisticRegression	PowerTransformer	√	0,8	0,37	0,58	0,27	0,54	0,61	0,8
LogisticRegression	QuantileTransformer	√	0,8	0,36	0,57	0,26	0,54	0,61	0,8

SHAP Values



Classifiers

- LogisticRegression
- RandomForest
- DecisionTree
- XGB
- LGBM
- GradientBoosting
- AdaBoost
- SVC

PCA

StandardScaler

SHAP Values

Stratify

SMOTE

White Wines

Classifier	Scaller	PCA	Stratify	Local Outlier	SMOTE	Score	Test F1 1	Test Precision 1	Test Recal 1	test balanced accuracy	test roc auc
RandomForest	Standard		√			0,9	**0,73**	**0,86**	0,64	0,64	**0,83**
RandomForest	Standard	√	√			0,89	**0,72**	**0,84**	0,63	0,63	**0,82**
RandomForest	Standard	√	√	√		0,88	0,66	**0,82**	0,56	0,62	**0,81**
RandomForest	Standard		√		√ 0,35 - (4898->4144)	0,9	0,74	**0,82**	0,67	0,64	**0,83**
RandomForest	Standard		√		√ 0,5 - (4898->4698)	0,89	0,73	**0,77**	0,7	0,65	**0,83**
RandomForest	Standard		√	√	√	0,87	0,67	0,74	**0,61**	**0,64**	**0,84**
SVC	Standard		√	√		0,78	0,62	0,5	**0,81**	**0,75**	**0,81**
RandomForest	Standard		√		√ 0,5 - (898->6140)	0,88	0.75	0.69	**0.81**	**0.6**	**0.84**
SVC	Standard	√	√	√		0,78	0,62	0,49	**0,83**	**0,74**	**0,81**
SVC	Standard	√	√			0,77	0,61	0,48	**0,84**	**0,74**	**0,81**
SVC	Standard		√			0,77	0,61	0,48	**0,84**	**0,74**	**0,81**

Red Wines

	Methodology	Score	F1 0	F1 1	Precision 0	Precision 1	Recal 0	Recal 1	best params	precision 1
0	GrindLogisticRegression	0.89	0.94	0.43	0.90	0.72	0.98	0.30	{'classifier__C': 4.0, 'classifier__solver': '...	0.72
1	GrindRandomForestClassifier	0.94	0.97	0.73	0.94	0.93	0.99	0.60	{'classifier__criterion': 'gini', 'classifier_...	0.93
2	GrindSVC	0.82	0.89	0.53	0.96	0.41	0.83	0.77	{'classifier__C': 2.0, 'classifier__gamma': 1....	0.41
0	GrindLGBMClassifier	0.93	0.96	0.73	0.95	0.82	0.98	0.65	{'classifier__boosting_type': 'dart', 'classif...	0.82
1	GrindXGBClassifier	0.95	0.97	0.80	0.96	0.82	0.97	0.77	{'classifier__booster': 'gbtree'}	0.82
2	GrindGradientBoostingClassifier	0.93	0.96	0.72	0.94	0.84	0.98	0.63	{'classifier__learning_rate': 1, 'classifier__...	0.84
3	GrindAdaBoostClassifier	0.90	0.94	0.54	0.92	0.70	0.97	0.44	{'classifier__algorithm': 'SAMME.R', 'classifi...	0.70

Sources

UCI Machine Learning Repository: Wine Quality Data Set

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

https://www.extension.iastate.edu/wine/total-sulfur-dioxide-why-it-matters-too/

Total Sulfur Dioxide – Why it Matters, Too!

Thank you for your attention!