Phase 2

Sebastian Sätzler, Bastian Tilk

Struktur

- Preprocessing
  - Skalierung
  - Outlier Removal
  - Standardisierung
- Modeling
  - Rotwein
  - Weißwein

Skalierung

MinMax Skalierung ausgewählt

- Über alle Features Skala [0;1]
- ...ausgenommen “quality”

→ RobustScaler nicht nötig, da Ausreißer im nächsten Schritt entfernt werden

Rotwein	
VIF:	
fixed acidity	53.426067
volatile acidity	16.703346
citric acid	9.108963
residual sugar	4.661322
chlorides	6.307859
free sulfur dioxide	6.356018
total sulfur dioxide	6.301839
density	4003.421361
pH	85.040860
sulphates	21.438226
magnesium	4.047249
alcohol	858.229937
lightness	1079.426466

Rotwein Skaliert	
VIF:	
fixed acidity	13.194483
volatile acidity	10.337223
citric acid	9.103255
residual sugar	2.636330
chlorides	5.241178
free sulfur dioxide	5.742105
total sulfur dioxide	5.284688
density	10.645272
pH	3.659068
sulphates	6.286883
magnesium	4.016390
alcohol	6.922062
lightness	18.308691

Outlier Removal

- Isolation Forest
- Local Outlier Factor
- Minimum Covariance Determinant
- One-Class SVM

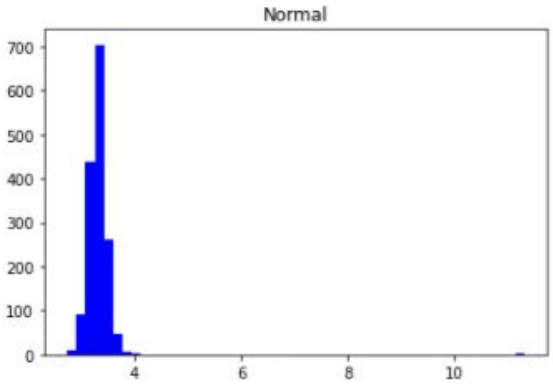
- Contamination auf 5% gesetzt
  - Anlehnung an Statistik mit 2 Standardabweichungen
  - Ergebnisse zufriedenstellend genug

- Validierung/Bewertung der Ergebnisse:
  - Beobachtung der Outlier Detection anhand Histogrammen von Features mit starken Outlier

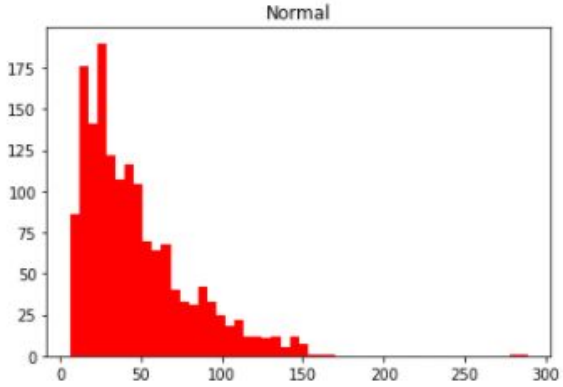
- Objektiv falsche Outlier manuell entfernt
  - pH von 99.99
  - Qualität von 99 und 17

Die betrachteten Features

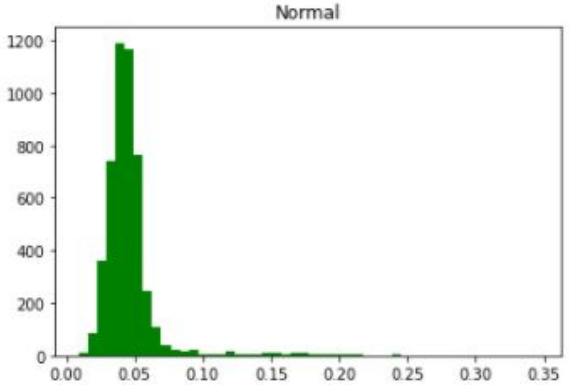
pH (Rotwein)



Total sulfur dioxide (Rotwein)



Chlorides (Weißwein)

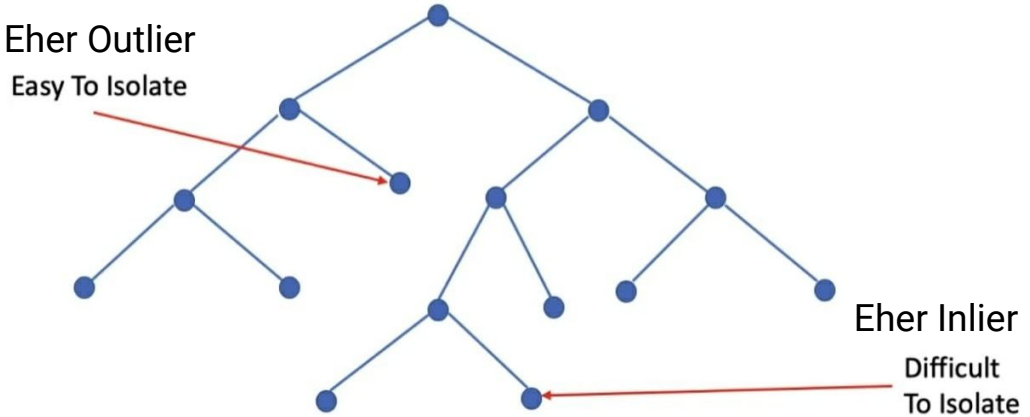


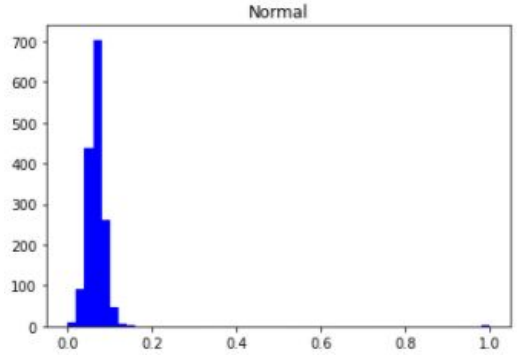
Isolation Forest

Generiert Random Forest Decision Tree

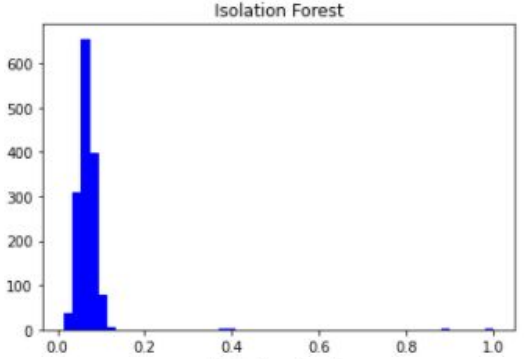
Häufig verwendet bei unsupervised learning
→ vielseitig anwendbar:

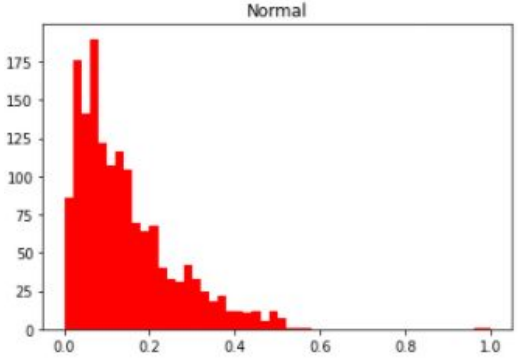
- Mit unterschiedlichen Skalenniveaus
- Anzahl der Dimensionen/Features
- Daten müssen nicht skaliert sein



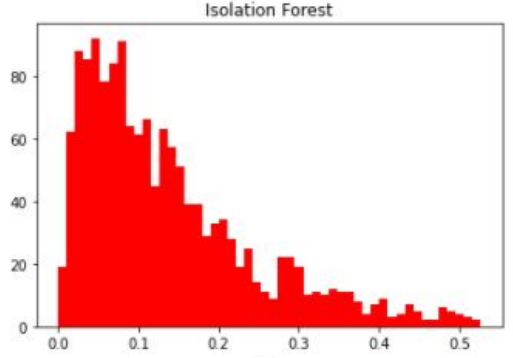


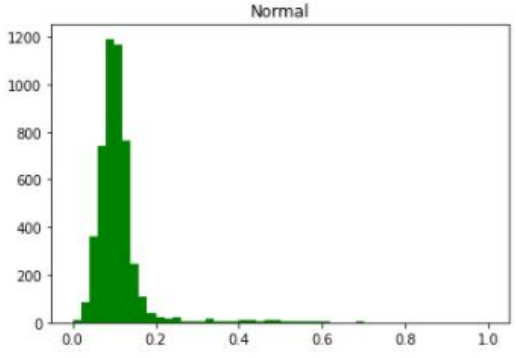
pH (Rotwein)



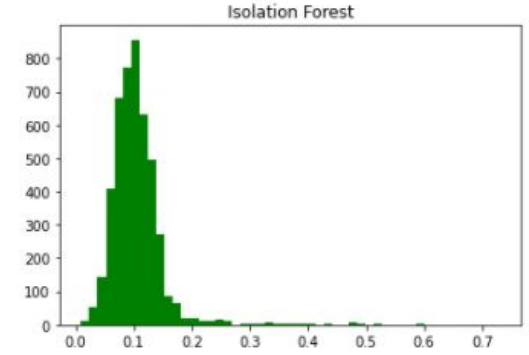


Total sulfur dioxide (Rotwein)



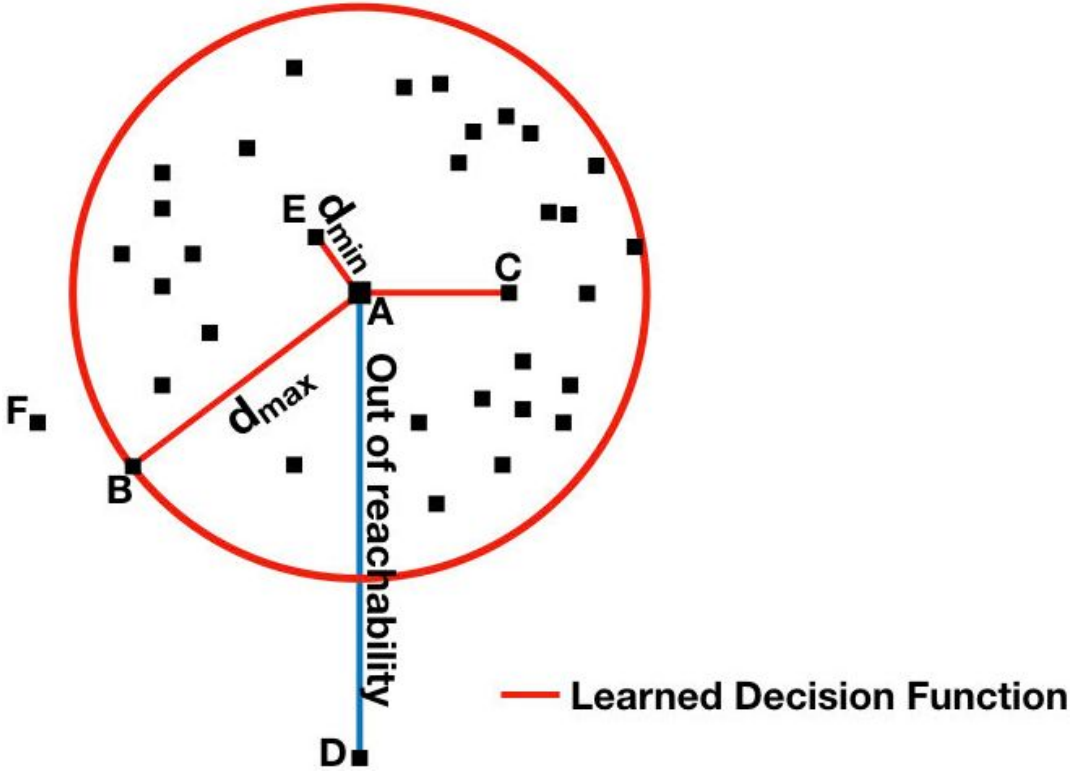


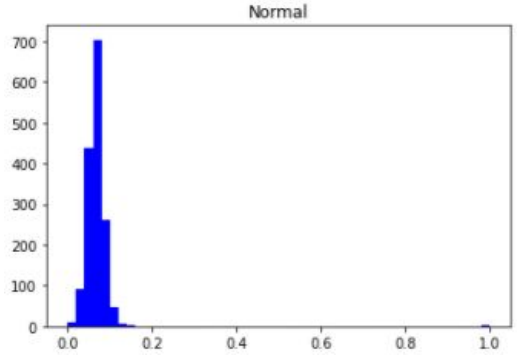
Chlorides (Weißwein)



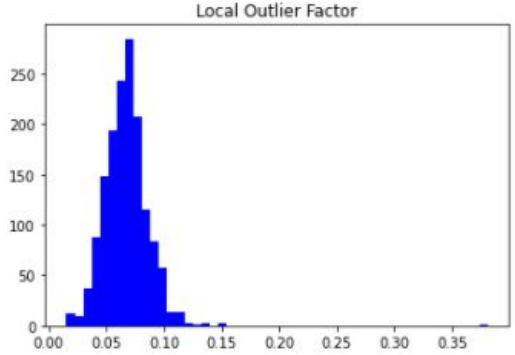
Local Outlier Factor

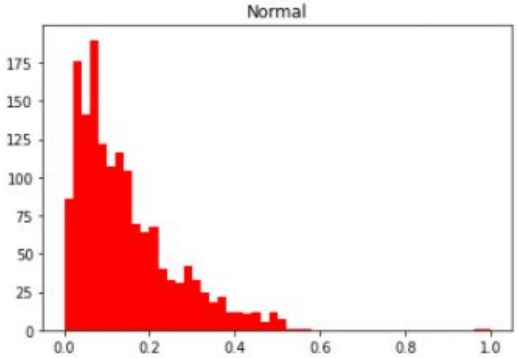
Density Based System
→ Skalierung des Datensatzes wichtig



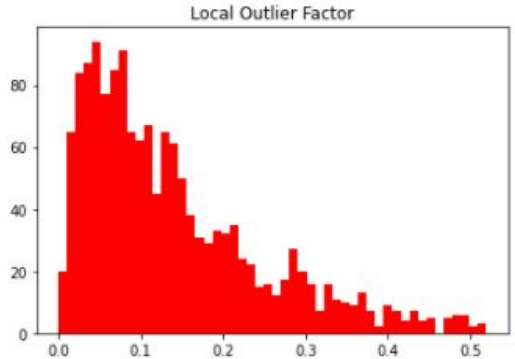


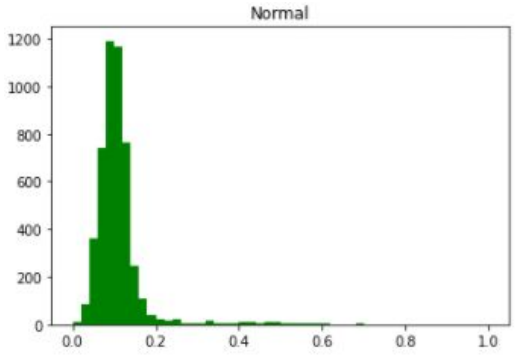
pH (Rotwein)



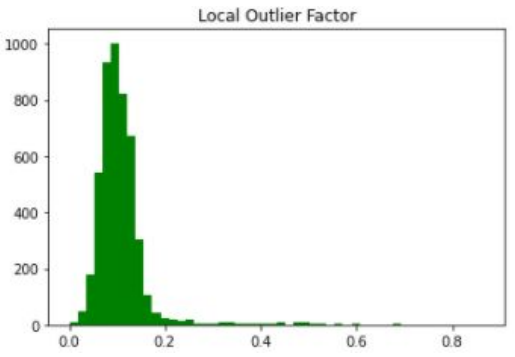


Total sulfur dioxide (Rotwein)





Chlorides (Weißwein)

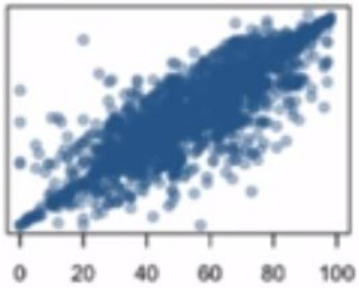
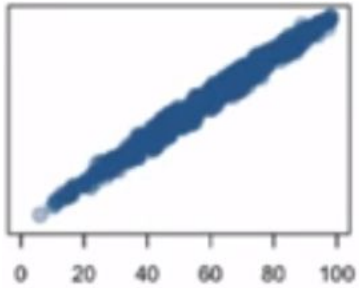


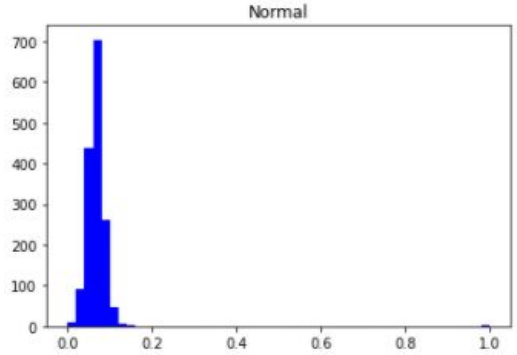
Minimum Covariance Determinant

Robust Estimator:
Wahre Zusammenhänge zwischen Features ausfindig machen

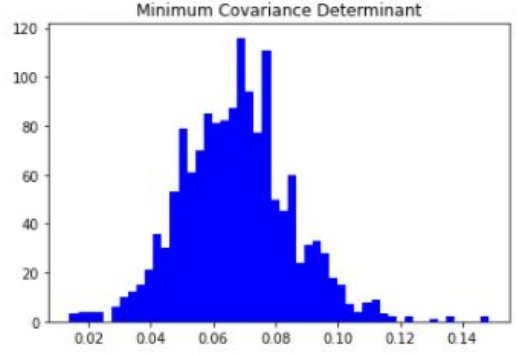
Berechnet die Korrelation, wie sie ohne Einfluss von Outlier wäre

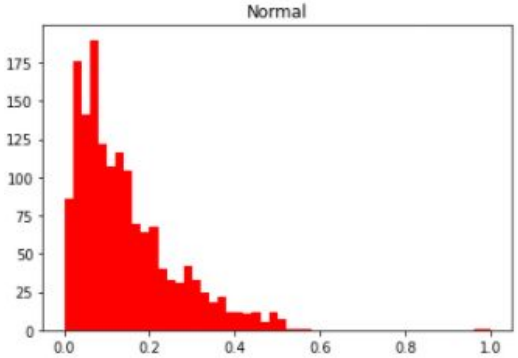
→ wird genutzt um Outlier zu entfernen

	Classical covariance	Robust covariance
		
Korrelation:	0.86	0.99

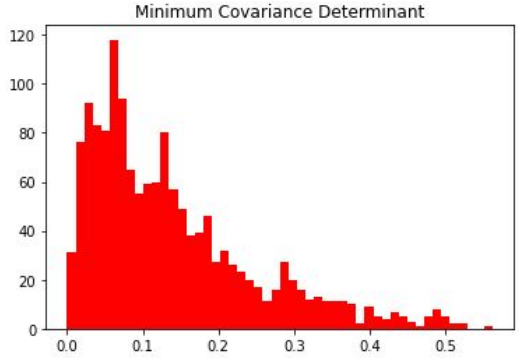


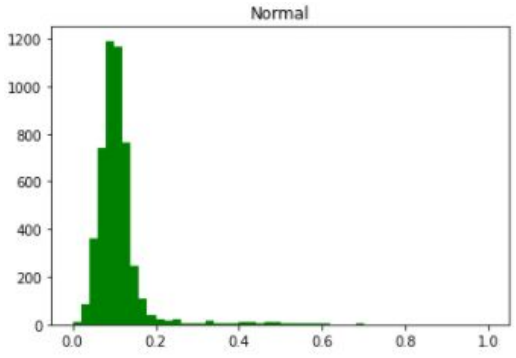
pH (Rotwein)



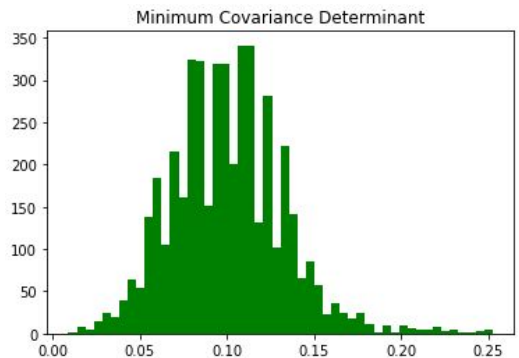


Total sulfur dioxide (Rotwein)





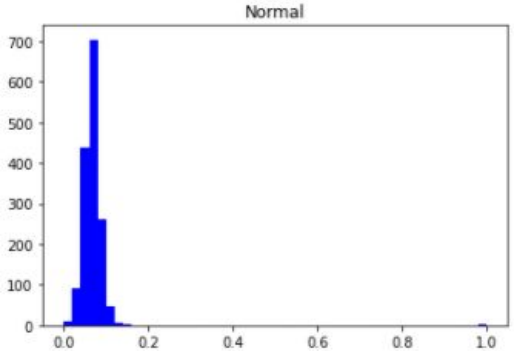
Chlorides (Weißwein)



Variance Influence Factor

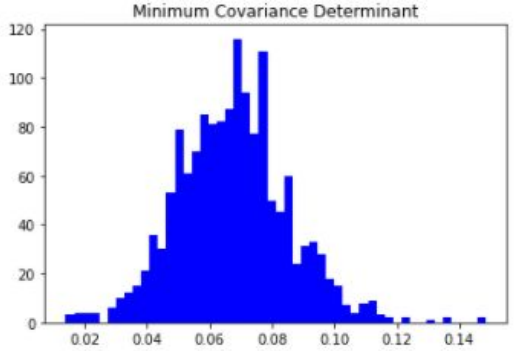
Rotwein Skaliert
VIF:

fixed acidity	13.194483
volatile acidity	10.337223
citric acid	9.103255
residual sugar	2.636330
chlorides	5.241178
free sulfur dioxide	5.742105
total sulfur dioxide	5.284688
density	18.645272
pH	3.659068
sulphates	3.286883
magnesium	4.016390
alcohol	6.922062
lightness	18.308691



Rotwein Skaliert, ohne Outlier
VIF:

fixed acidity	19.130504
volatile acidity	10.933478
citric acid	9.068865
residual sugar	4.857854
chlorides	14.750573
free sulfur dioxide	6.121432
total sulfur dioxide	5.698108
density	11.472832
pH	21.263061
sulphates	7.541263
magnesium	4.001089
alcohol	14.367848
lightness	26.703889



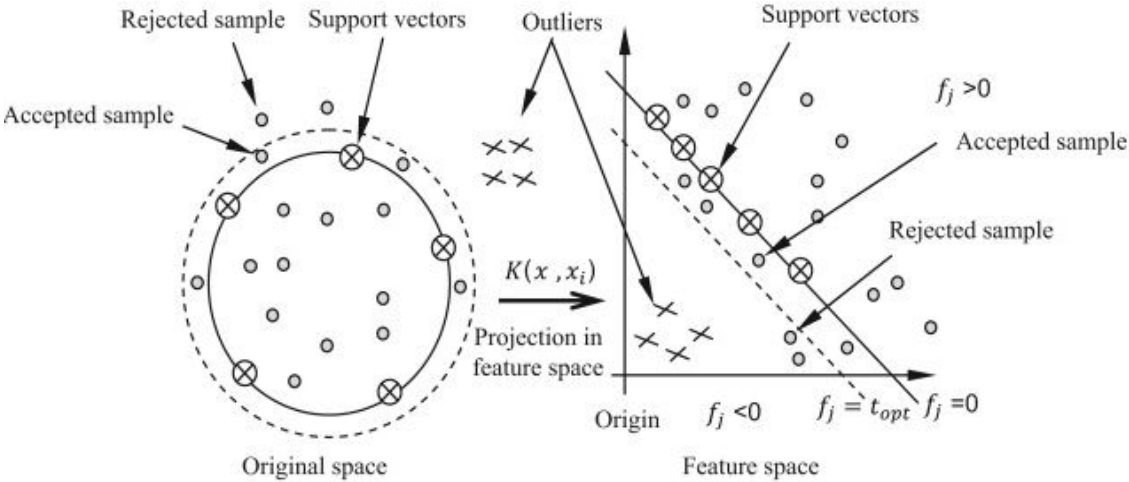
One-Class SVM

Novelty Detection
→ Daten die selten auftauchen/Outlier

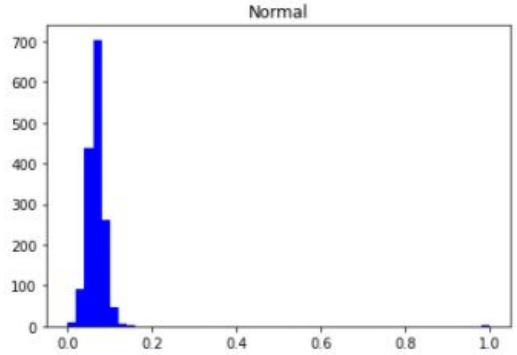
Hypersphere wird angelegt

Durch Feature expansion linear separierbar (SVM anwenden)
- sehr aufwendig v.a. mit hochdimensionalen Daten (viele Features) O(n^2)

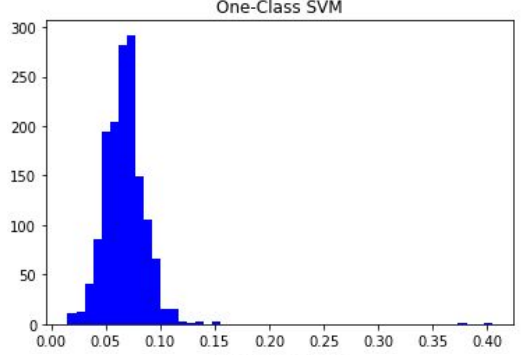
⇒ Verwendung des Kernel-Trick
- ermöglicht Klassifizierung, ohne dabei im Feature space zu arbeiten

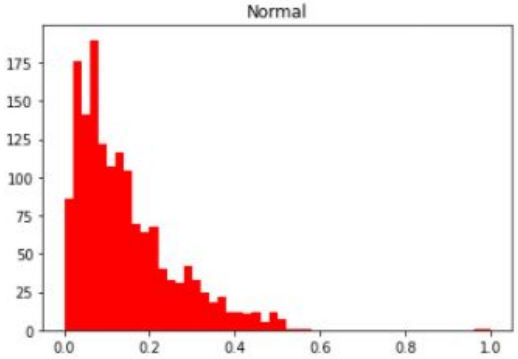


Feature expansion

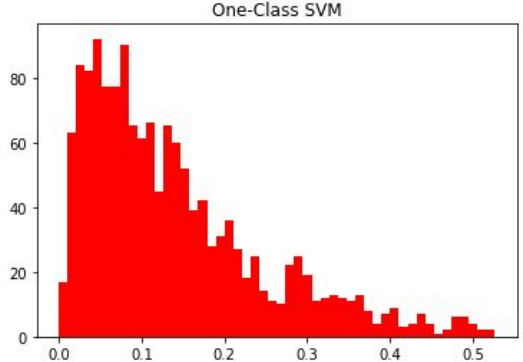


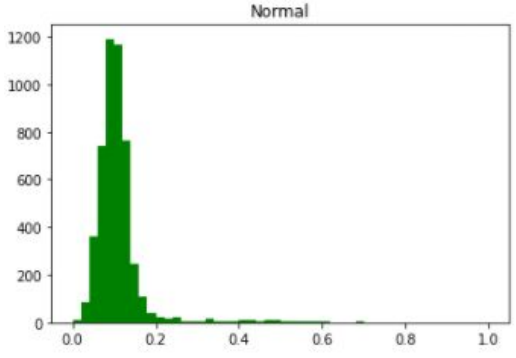
pH (Rotwein)



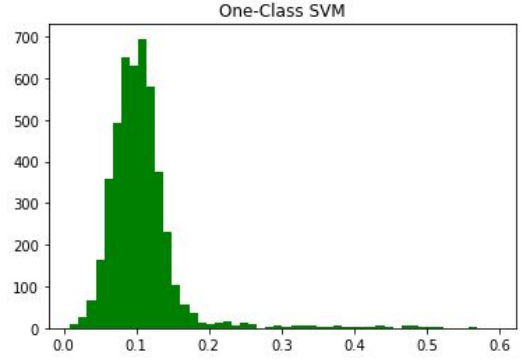


Total sulfur dioxide (Rotwein)



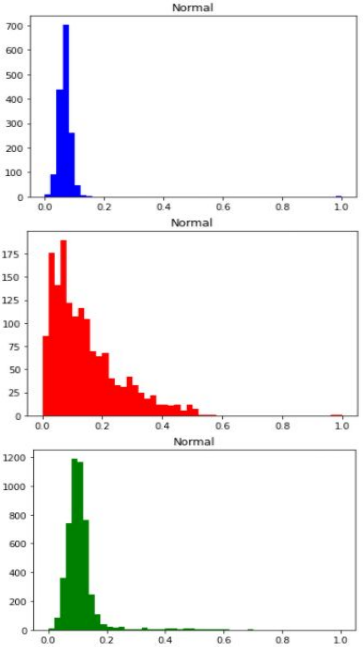


Chlorides (Weißwein)

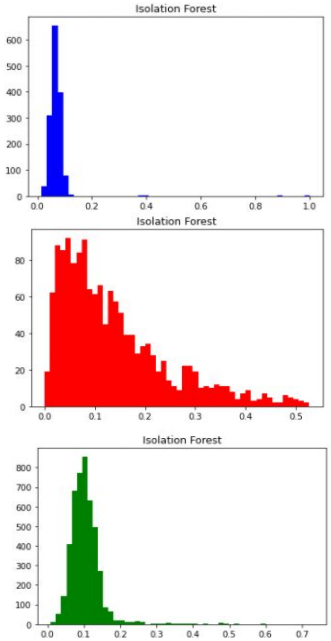


Überblick

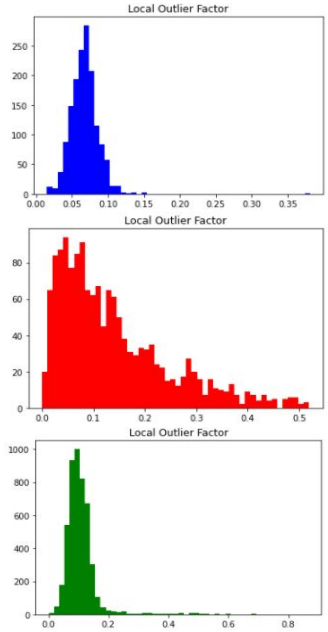
Kompletter Datensatz



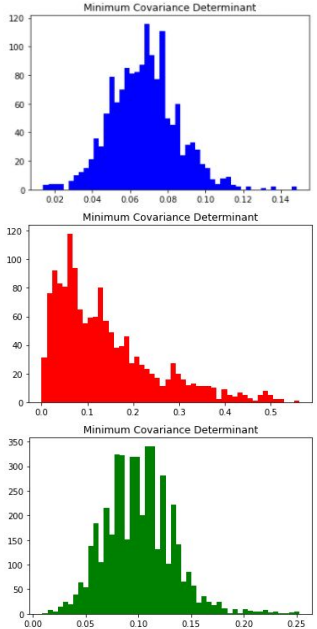
Isolation Forest



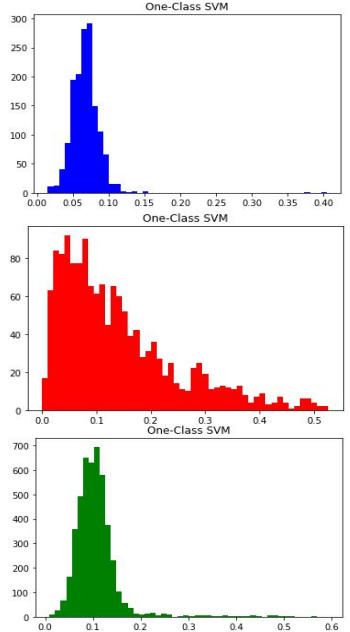
Local Outlier Factor



Minimum Covariance Determinant



One-Class SVM



Überblick

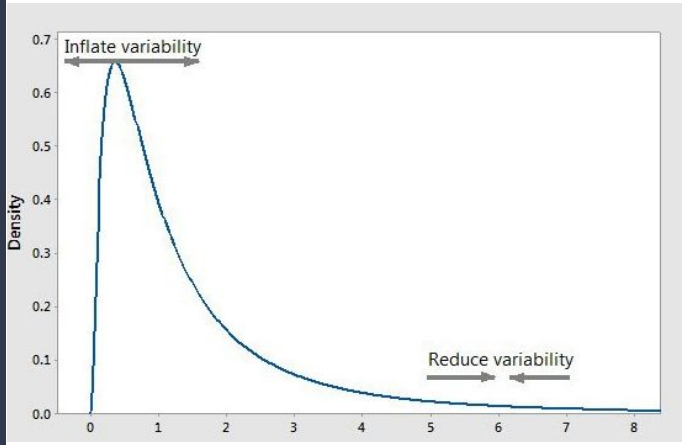
	Kompletter Datensatz	Isolation Forest	Local Outlier Factor	Minimum Covariance Determinant	One-Class SVM
Weißwein	4881	4637	4843	4636	4640
Rotwein	1564	1485	1518	1485	1484

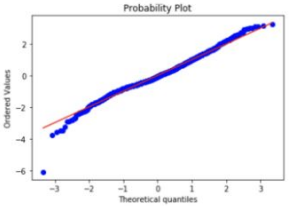
Standardisierung-PowerTransformer

Ziemlich effektiv, um Datensätze zu Standardisieren → Erwartungswert 0, Varianz 1 (Gauss Glockenkurve)

Daten verlieren an lesbarkeit (für den Menschen)...

... Aber das spielt bei der Modellerstellung eine geringere Rolle, solange die Accuracy besser wird





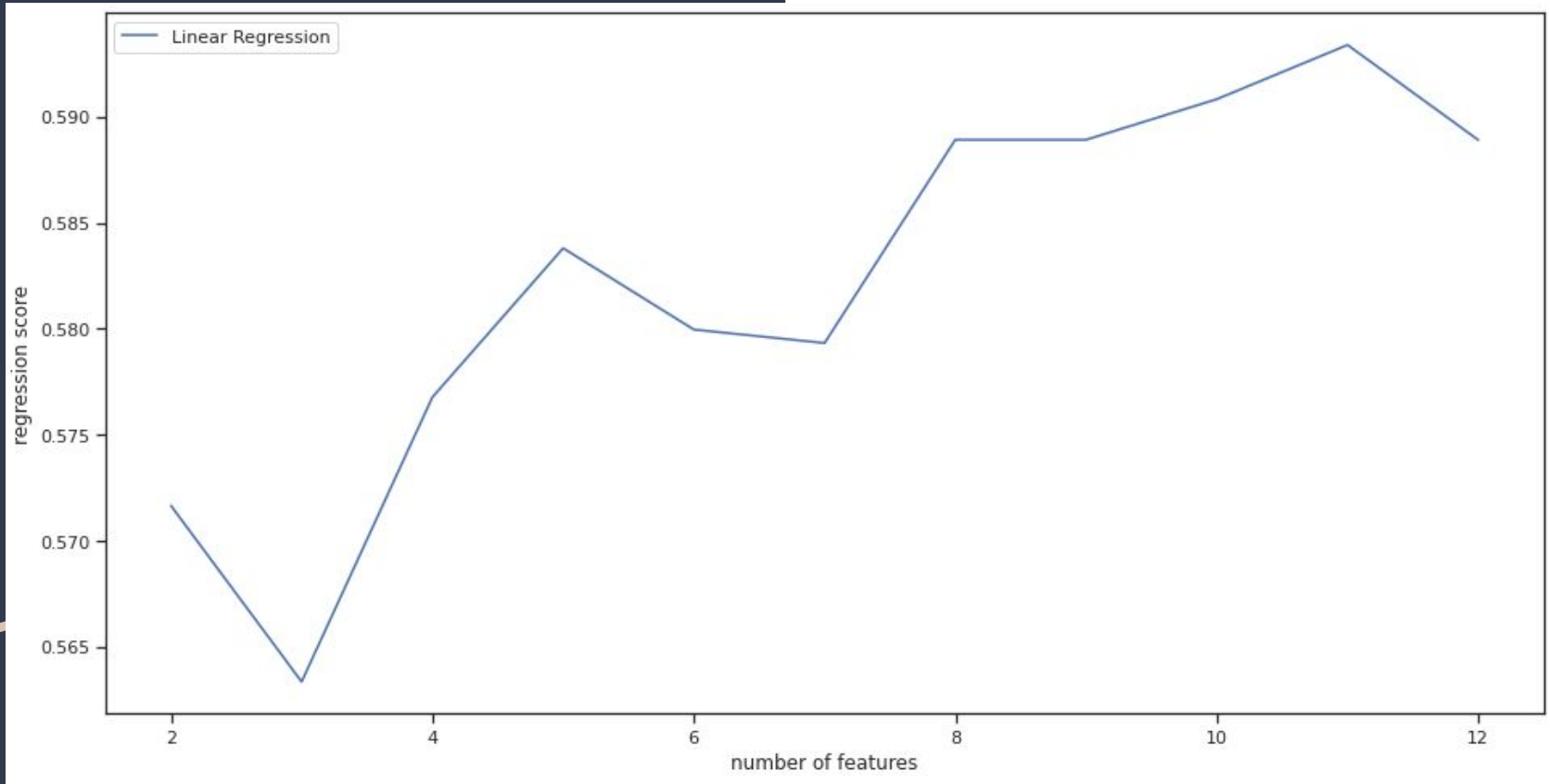
Modellerstellung

Manuelles Preprocessing

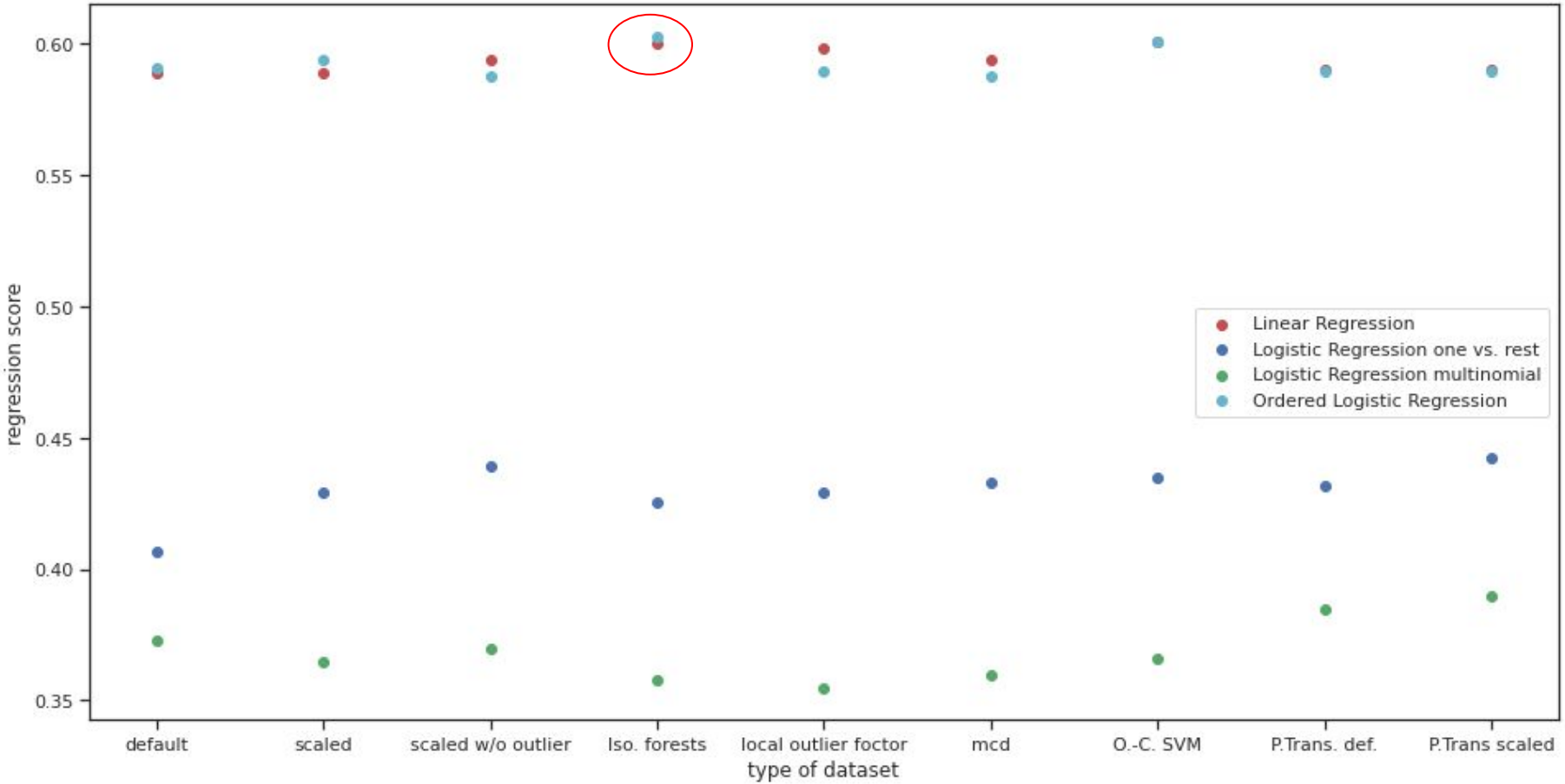
Allgemein:
- Datensätze mit Missing Values entfernt
- Noise bei quality und pH-Wert entfernt
- Feature flavanoids entfernt

Bei Rotwein:
- density entfernt

Rotwein Standard Datensatz mit Information Gain Classifier



Rotwein Regressionsmethoden und Datensätze

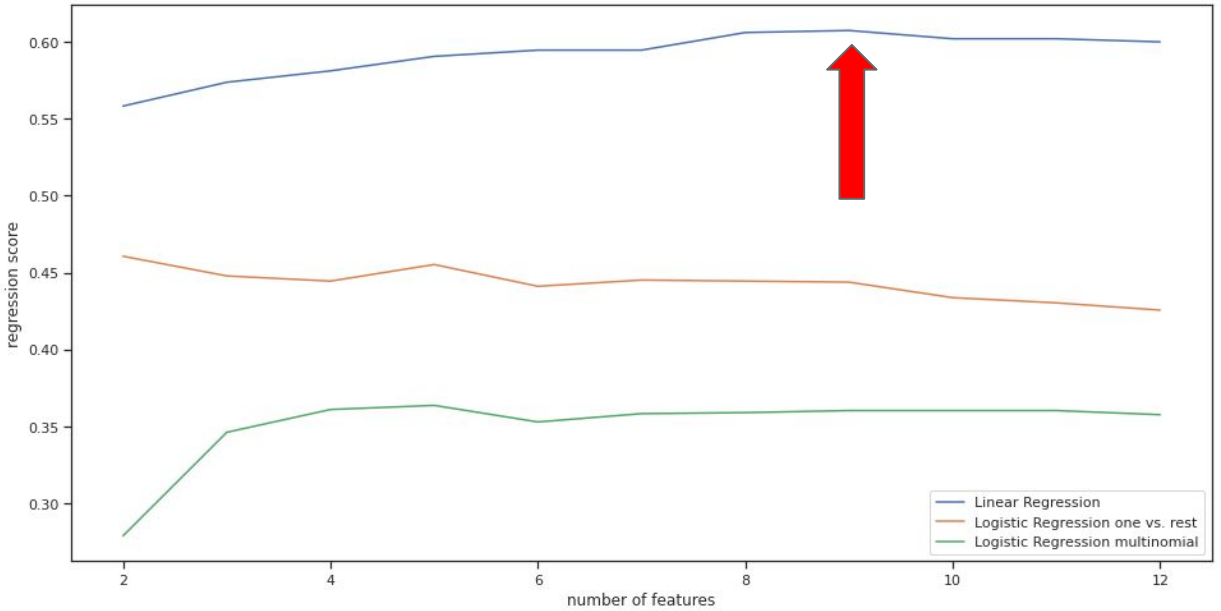


Rotwein: Accuracy erhöhen

Information Gain Classifier mit Datensatz der durch Isolation Forests von Outlier befreit wurde.

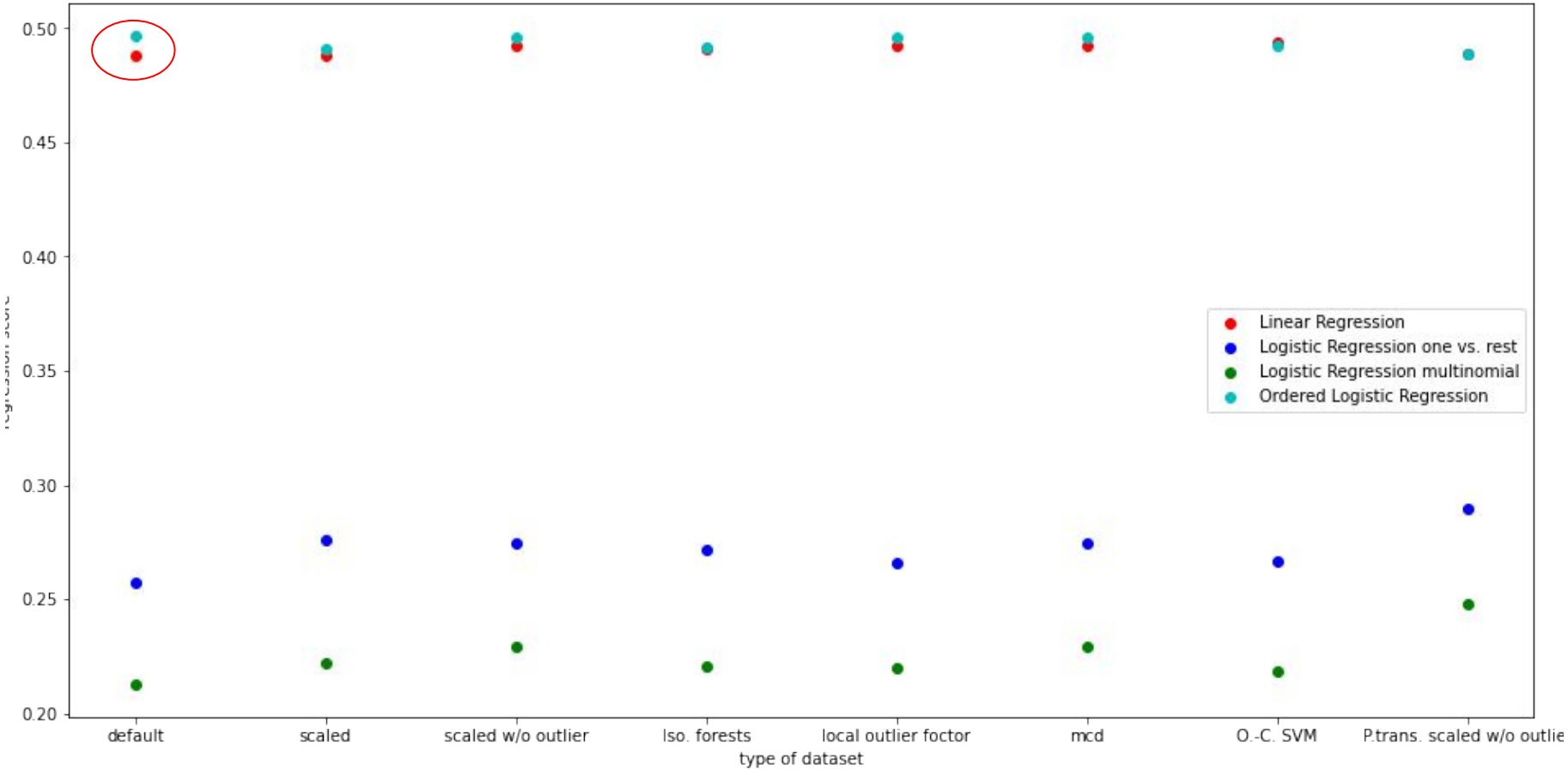
Höchste Genauigkeit mit 9 Features:
60,94%

- Fixed acidity
- Volatile acidity
- Citric acid
- Chlorides
- Total sulfur dioxide
- Sulphates
- Magnesium
- Alcohol
- lightness

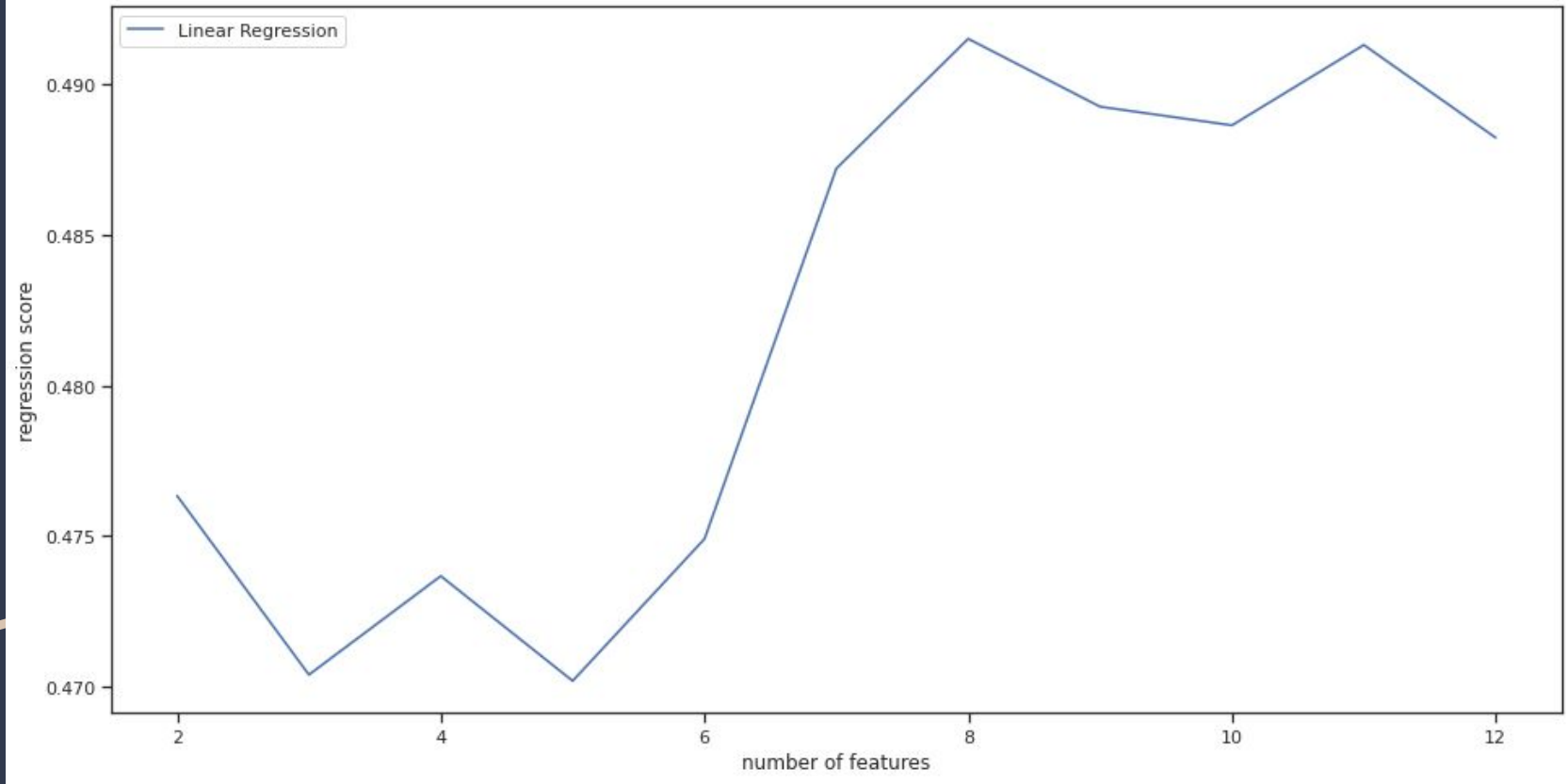


Weißwein

Weißwein Regressionsmethoden und Datensätze



Weißwein Standard Datensatz mit Information Gain Classifier (IGC)



Weißwein Standard Datensatz

Höchste Genauigkeit im IGC von 49,1% mit 8 Features:

- residual sugar
- Chlorides
- free sulfur dioxide
- total sulfur dioxide
- Density
- Magnesium
- Alcohol
- Lightness
- quality

Vgl. Korrelationstabelle aus Phase 1

Pearson Spearman

Korrelation	Weißwein	Rotwein
stark	residual sugar & density (0.84) alcohol & lightness (-0.86)	alcohol & lightness (-0.95)
moderat	free sulfur dioxide & total sulfur dioxide (0.62) density & lightness (0.69) density & alcohol (-0.78)	fixed acidity & citric acid (0.67) free sulfur dioxide & total sulfur dioxide (0.67) fixed acidity & pH (-0.71) citric acid & pH (-0.57)
schwach	total sulfur dioxide & residual sugar (0.4) total sulfur dioxide & density (0.53) total sulfur dioxide & lightness (0.4) residual sugar & lightness (0.42) residual sugar & alcohol (-0.45) total sulfur dioxide & alcohol (-0.45) chlorides & density (0.51) chlorides & lightness (0.5) chlorides & alcohol (-0.57)	density & lightness (0.48) alcohol & quality (0.47) volatile acidity & citric acid (-0.56) density & alcohol (-0.52) lightness & quality (-0.44)

Weißwein Optimierung (1)

Höchste Genauigkeit von 49,97% (+0,2%) mit 11 Features:

- fixed acidity
- volatile acidity
- citric acid
- residual sugar
- chlorides
- free sulfur dioxide
- total sulfur dioxide
- pH
- sulphates
- magnesium
- alcohol

Weißwein Optimierung (2)

Formatierung von Features mit hoher Varianz/Uniformverteilung:

- total sulfur dioxide
- free sulfur dioxide

Genauigkeit: 49,1 (-0,8%)

Entfernung von Features mit hoher Korrelation:

- lightness
- density
- citric acid
- fixed acidity
- total sulfur dioxide

Genauigkeit: 49,1 (-0,8%)

Ergebnis

- Rotwein 60,9% (Lineare Regression)
- Weißwein 49,9% (Lineare Regression)

Fazit

- Ergebnisse nicht Vorhersehbar
  - Prognosen beim Preprocessing haben sich teilweise nicht bewahrheitet
  - Iteratives/Experimentelles Vorgehen essentiell! → Ausprobieren

- Rotwein liefert ein besseres Modell als Weißwein
  - Bereits in Task 1a ersichtlich:
    → Rotwein hat korrelierende Features mit dem Qualität-Attribut