**Projektbericht Data Mining**

Prüfer: Prof. Dr. Schicker

12.01.2022

Stefan Butz, [stefan.butz@st.oth-regensburg.de](mailto:stefan.butz@st.oth-regensburg.de), Matrikelnummer: 3175600

Ronny Pollak, [ronny.pollak@st.oth-regensburg.de](mailto:ronny.pollak@st.oth-regensburg.de), Matrikelnummer: 3173031

Inhaltsverzeichnis

[Einleitung 3](#_Toc92862364)

[Vorstellung des Datensatzes 3](#_Toc92862365)

[Aufbau 3](#_Toc92862366)

[Bereinigung 3](#_Toc92862367)

[Regressionsanalyse 5](#_Toc92862368)

[Ziel 5](#_Toc92862369)

[Vorgehen 6](#_Toc92862370)

[Erkenntnisse und Aussicht 10](#_Toc92862371)

[Klassifikationsanalyse 10](#_Toc92862372)

[Ziel 10](#_Toc92862373)

[Vorgehen 10](#_Toc92862374)

[Erkenntnisse 13](#_Toc92862375)

[Segmentierung 13](#_Toc92862376)

[Assoziationsanalyse 13](#_Toc92862377)

[Verwendete Verfahren 14](#_Toc92862378)

[Chemische Zusammenhänge 14](#_Toc92862379)

[Säureanteile - pH-Wert 14](#_Toc92862380)

[Säureanteile - Alkohol – Dichte 16](#_Toc92862381)

[Schwefeldioxide 17](#_Toc92862382)

[Qualitative Zusammenhänge 18](#_Toc92862383)

[Qualitätsfördernde Eigenschaften 18](#_Toc92862384)

[Erkenntnisse 19](#_Toc92862385)

[Fazit 19](#_Toc92862386)

[Abbildungsverzeichnis 20](#_Toc92862387)

[Tabellenverzeichnis 20](#_Toc92862388)

[Literaturverzeichnis 20](#_Toc92862389)

# Einleitung

Portugal zählt mit Volumen von 3.1 Millionen Hektolitern Wein seit zu den zehn größten Weinexporteuren im Jahr 2020 [1]. Einer der bekanntesten portugiesischen Weine ist der vinho verde. Die offizielle Zertifizierungsstelle für vinho verde (CVRVV) hat die Aufgabe die Qualität der Weine sicherzustellen. Zur Bestimmung der Qualität werden verschiedene chemische Kennzahlen erfasst und Geschmacksproben durchgeführt [2]. Der Fortschritt in der Informationstechnologie erlaubt es alle ermittelten Kennwerte zu speichern und mithilfe von Data Mining Werkzeugen zu analysieren. Im Folgenden wird versucht die Qualität des Weines auf Basis der chemischen Kennzahlen zu bestimmen.

# Vorstellung des Datensatzes

## Aufbau

Der Datensatz wurde in einer Fallstudie von Paul Cortez in Zusammenarbeit mit dem CVRVV erstellt und beinhaltet 4898 rote und 1599 weiße Weinproben der Sorte vinho verde aus der Region Minho. Eine Probe bezieht sich immer auf genau einen Wein. Pro Probe enthält der Datensatz die Ergebnisse von elf analytischen Tests und die sensorische Bewertung des Weins. Die sensorische Bewertung wurde von mindestens drei Bewertern in Blindtests bestimmt [2].

Die Weinproben liegen nach Traubenfarbe getrennt in zwei CSV-Dateien vor. Diese haben beide den gleichen Aufbau und haben folgende Spalten:

1. fixed acidity: Fester Säuregehalt in g/L
2. volatile acidity: Flüchtiger Säuregehalt in g/L
3. citric acid: Citronensäure in g/L
4. residual sugar: Restzucker in g/L
5. chlorides: Chloride in g/L
6. free sulfur dioxide: Freie Schwefeloxide in mg/L
7. total sulfur dioxide: Schwefeloxide in mg/L
8. density: Dichte in g/cm3
9. pH: pH-Wert
10. sulphates: Sulfate in g/L,
11. alcohol: Alkohol in vol.%
12. quality: Qualität 1 (=sehr schlecht) bis 10 (=sehr gut)

Die ersten elf Spalten sind numerische Werte. Die letzte Spalte weist nur ganzzahlige Werte zwischen 1 und 10 auf. Da die Werte zwischen 1 und 10 nicht nur eine Klassenzugehörigkeit angeben, sondern auch eine Ordnung besitzen, kann die letzte Spalte auch als numerischer Wert betrachtet werden.

## Bereinigung

Der Datensatz beinhaltet keine Null-Werte. Eine Entfernung oder Ersetzung von Null-Werten ist also nicht notwendig.

Abbildung 1 ist ein Box-Plot der Rotweine. Abbildung 2 ist ein Boxplot der Weißweine. Die Abbildungen bilden pro Attribut das gesamte Wertespektrum ab. Ausreißer werden als Kreise in den rechten und linken äußeren Bereichen dargestellt. Auf den Abbildungen sieht man, dass es in fast allen Spalten Ausreißer gibt. Um eine Ausgabe aller Ausreißer zu ermöglichen, werden alle Proben, die in einer Spalte um mehr als 2,5 Standardabweichungen vom Durchschnitt der jeweiligen Kennzahl abweichen, als Ausreißer markiert. Eine genaue Untersuchung der markierten Ausreißer zeigt, dass selbst diese die in Kapitel Chemische Zusammenhänge beschriebenen Zusammenhänge aufweisen. Es handelt sich daher bei den Ausreißern nicht um Messfehler. Die Ausreißer könnten Nischenprodukte, Fehlprodukte oder eine Folge natürlicher Schwankungen bei Naturprodukten wie Wein sein.

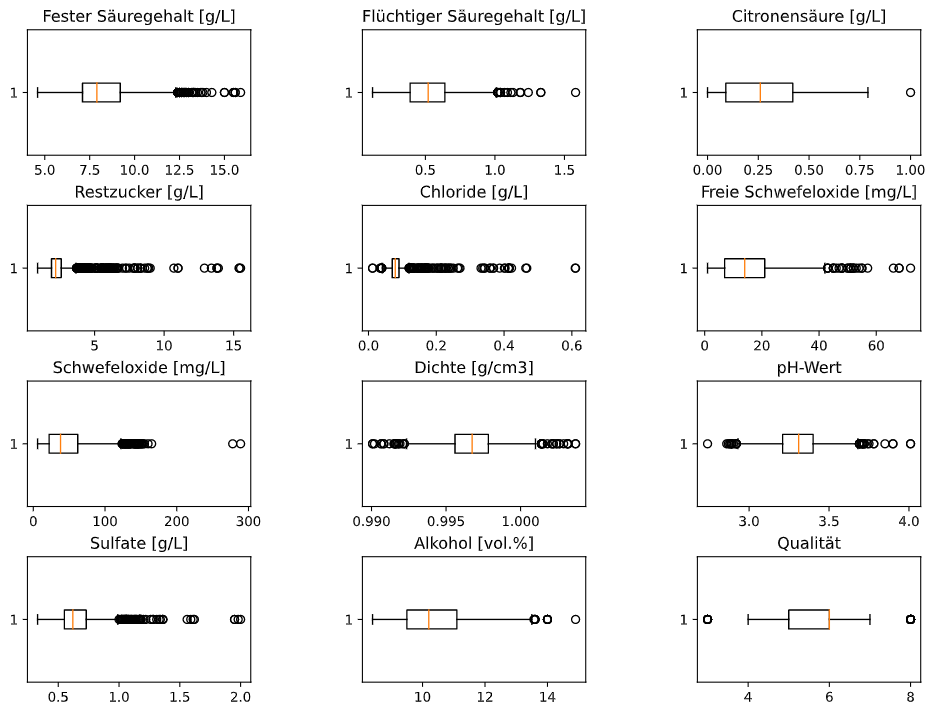


Abbildung 1: Box-Plot aller Attribute der Rotweine

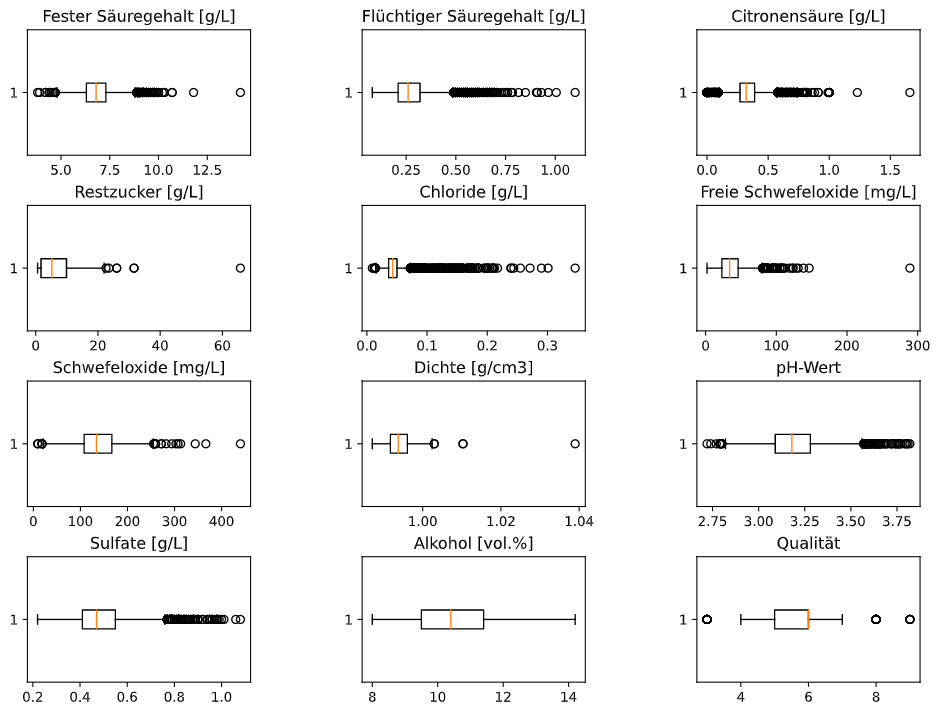


Abbildung 2:Box-Plot aller Attribute der Weißweine

Da die Qualitätsbestimmung von Nischenprodukten kein Ziel dieser Arbeit ist und die Verkleinerung des Wertebereichs sich positiv auf die Fähigkeiten von Datamining-Verfahren, wie bspw. Neuronale Netze auswirkt, werden die Ausreißer entfernt. Ausgenommen davon werden Ausreißer der Spalte Qualität, da sich sonst das Wertespektrum der Qualität verändern würde.

Die nachfolgenden Analysen werden, falls nicht anders vermerkt, stets mit den gefilterten Daten durchgeführt. Der Fokus liegt stets auf den Rotweinen, da der Datensatz der Weißweine den vorgesehenen Umfang des Projekts übersteigt.

# Regressionsanalyse

## Ziel

Zum Start des Projektes wurde die Vorhersage einer Spalte mithilfe der Regressionsanalyse als Ziel genommen. Es sollte die Qualität mithilfe der anderen elf Spalten geschätzt werden.

## Vorgehen

Zu Anfang wurden die beiden Datasets (Rotwein und Weißwein) eingelesen und in einem Dataframe gespeichert. Diese beiden Dataframes wurden dann zu einem weiteren großen Dataframe zusammengefügt, der die Daten beider beinhaltet. Jedem dieser drei Dataframes wurde zusätzlich eine Beschreibung hinzugefügt, um diese in der folgenden Analyse besser zu unterscheiden. Mittels einer Schleife wurde nun das nachfolgende Prozedere auf jeden der drei Dataframes angewandt.

Zuerst wurde eine Scattermatrix hinsichtlich der X-Spalten und der zu vorhersagenden Y-Spalte, der Qualität, erzeugt, um mögliche Zusammenhänge zu erkennen.

Ein Bild, das Text, Vorhang, Gewebe enthält.

Automatisch generierte Beschreibung

Abbildung 3: Scattermatrix Rotwein

Wie in Abbildung 3 zu sehen ist, kann man auf den ersten Blick keine linearen Zusammenhänge zwischen der Qualität und den übrigen Spalten erkennen. Somit wurden die folgenden Methoden der Regression mithilfe aller Spalten durchgeführt. Zuerst wurde das jeweilige Dataframe in eine X und y Komponente, anhand der Spalten, aufgeteilt und mittels train\_test\_split des Packages sklearn in Trainings- und Testdaten geteilt. Hierfür wurde eine Testgröße von 30% verwendet. Der Zufallsstrom wurde mehrmals geändert um Abweichungen zu erkennen, führte jedoch zu sehr ähnlichen Ergebnissen in der Analyse.

Als erstes Modell wurde LinearRegression aus dem Paket sklearn verwendet. Dieses Modell wurde, wie typisch für eine Regression, anhand des R2-Wertes bewertet. Dies führte zu sehr schlechten Ergebnissen. In Abbildung 4 werden die R2-Werte für Trainings- und Testdaten des Datensatzes des Rotweins erkenntlich. Das Ergebnis des R2-Wertes des Weißweins und der beiden Datensätze zusammen war um jeweils 0.07-0.08 für beide Werte schlechter.



Abbildung 4: Ergebnisse der linearen Regression bei Rotweinen

Der nächste Versuch wurde mit einer Polynomialen Regression durchgeführt. Die verwendete Methode war wieder LinearRegression, jedoch wurden zuvor Trainings- und Testdaten polynomiell transformiert. Diese Regression wurde sowohl auf Grad 2 als auch 3 durchgeführt.

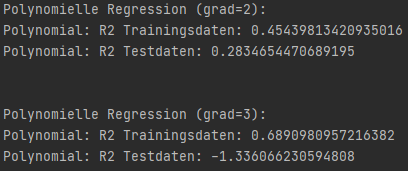


Abbildung 5: Ergebnisse der polynomiellen Regression bei Rotwein

Wie in Abbildung 5 zu sehen, wurden somit die Ergebnisse der Trainingsdaten des Rotweins verbessert, jedoch die der Testdaten verschlechtert, und bei Grad 3 sogar ins Negative befördert. Die Daten der beiden weiteren Datensätze (Weiß und Gemischt) machten eine sehr ähnliche Entwicklung. Dies deutet stark auf Overfitting, also eine zu starke Anpassung an die Trainingsdaten, hin und macht diese Methode somit nutzlos.

Die nächste verwendete Methode war die Bayessche lineare Regression. Hier wurde der BayesianRidge Algorithmus mit 300 Iterationen und einer Toleranz von 0,001 angewendet. Die Ergebnisse für alle drei Datensätze ähnelten hier jedoch wieder stark der zu Anfang durchgeführten linearen Regression. Dadurch wurden keine Verbesserungen oder neuen Erkenntnisse gewonnen.

Daraufhin wurde eine Support Vector Regression durchgeführt. Hierfür wurde die Methode des Packages sklearn benutzt. Die Methode wurde jeweils mit einem der vier verschiedenen Kernels linear, poly, rbf und sigmoid aufgerufen.

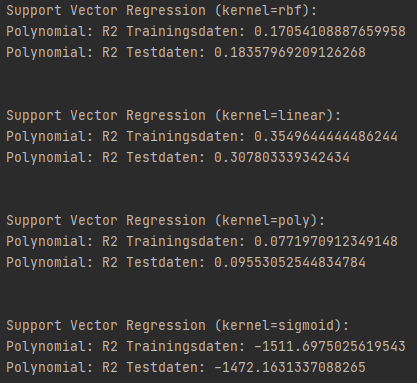


Abbildung 6: Ergebnisse der Support-Vector-Regression bei Rotweinen

Wie in Abbildung 6 zu sehen ist brachten diese Berechnungen jedoch wieder nur schlechtere Ergebnisse. Das beste Ergebnis, die lineare Support Vector Regression, nähert sich der zu Anfang durchgeführten linearen Regression an, welche jedoch auch ein schlechtes Ergebnis zurückgibt.

Ein weiterer Versuch wurde mit einem Entscheidungsbaum unternommen. Die hierfür verwendete sklearn Methode DecisionTreeRegressor wurde mit den vier verschiedenen Kriterien, welche die Funktion, die die Qualität des Splits bewertet, darstellt, getestet. Des Weiteren wurde als Splitverhalten das beste (weitere Option wäre zufällig), als maximale Tiefe keine und als minimale Anzahl an Elementen für einen Split 1% festgelegt.

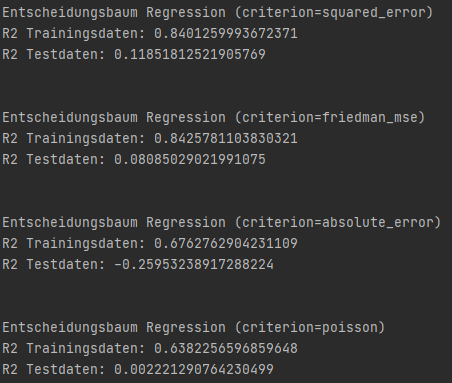


Abbildung 7: Ergebnisse der Entscheidungsbaum-Regression bei Rotweinen

Die R2-Werte in Abbildung 7 zeigen mittlere bis gute Ergebnisse für Trainingsdaten, jedoch schlechte Ergebnisse für Testdaten, was für einen Entscheidungsbaum nicht unüblich ist und auf Overfitting deutet.

Für den letzten Regressionsversuch wurde ein Neuronales Netz verwendet. Hier wurde ein sequentielles Neuronales Netz mit drei Schichten aufgebaut. Die erste Schicht besteht aus zwölf Neuronen, der Input-Dimension elf (Anzahl der X Parameter) und der Aktivierungsfunktion relu. Die zweite Schicht besteht aus 100 Neuronen und ebenfalls der Aktivierungsfunktion relu. Die letzte Schicht besteht wieder einem Neuron und der Aktivierungsfunktion linear.

Dieses Modell wurde dann mit dem Compiler Adam, dem loss als mean squared error und der Metrik als mean absolute error kompiliert. Das Neuronale Netz wurde im Anschluss mit 100 Epochen und denselben Test und Trainingsdaten der vorherigen Regressionen trainiert. Wie man in Abbildung 8 sehen kann, war auch dieses Modell nicht erfolgreich.

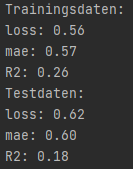


Abbildung 8: Ergebnisse der Regression mit neuronalen Netzen bei Rotweinen

Die Daten in Abbildung 9 zeigen, dass die Loss-Kurve und der mean absolute error konvergieren, woraus man schließen kann, dass dieses Ergebnis nicht das Resultat einer zu geringen Anzahl an Epochen ist.

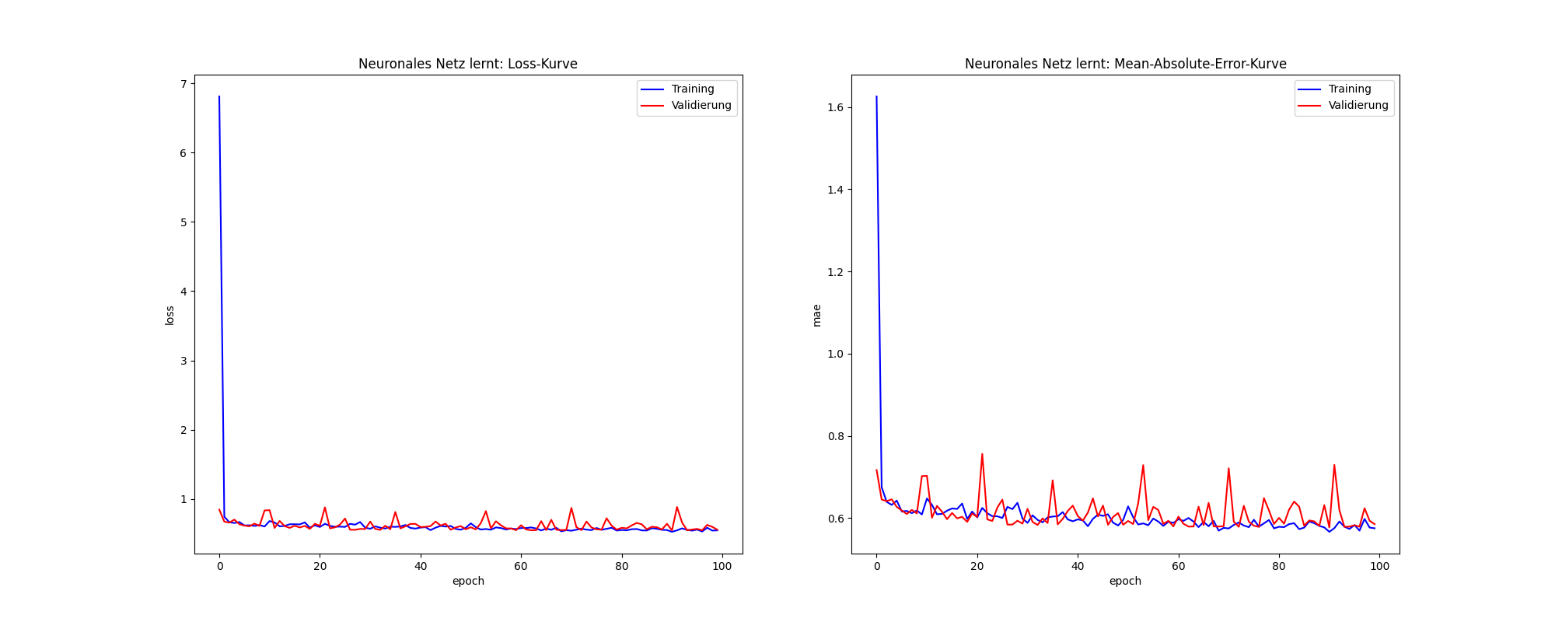


Abbildung 9: Lernkurve der Neuronalen Netze bei Regression von Rotweinen

## Erkenntnisse und Aussicht

Wie man in den zu vorgehenden Beispielen sehen kann, war die Regressionsanalyse nicht erfolgreich. Die einzigen guten Werte wurden durch Overfitting erzielt, welches das Modell für Testdaten unbrauchbar macht. Des Weiteren sind die Ergebnisse jedoch stabil. Durch eine Änderung von Zufallsströmen bzw. Datensätzen (Weiß, Rot, Gesamt) gibt es keine signifikanten Änderungen im Ergebnis.

Um die Ergebnisse zu verbessern könnte man die Prognose toleranter gestalten, d.h. wenn ein vorhergesagter Wert nahe an seinem realen Wert ist, wird dieser als korrekt gewertet. Wenn zum Beispiel die Qualität eines Weines als 7 prognostiziert wird, er jedoch in Wahrheit 8 ist, wird dieser als richtig gewertet (+-1).

Eine weitere Möglichkeit ist die Einteilung in Klassen, das heißt, dass der Wein in verschiedene Qualitätsstufen unterschieden wird. In der folgenden Klassifikationsanalyse wird dieses Verfahren angewandt.

# Klassifikationsanalyse

## Ziel

Das Ziel der Klassifikation war es, die Weine anhand ihrer Qualität in Qualitätsstufen einzuteilen und einem Modell diese Daten dann den einzelnen Klassen zuzuordnen lassen. Es wurden wieder dieselben Datensätze wie zuvor benutzt und es werden die Ergebnisse des Rotwein Datensatzes präsentiert.

## Vorgehen

Zuerst wurden die Daten in ihren Original-Qualitätsklassen (1-10) belassen und versucht diese als Klassifikationsgruppen zu verwenden. Dafür wurde das Naive Bayes Modell verwendet und mit Trainings- und Testdaten gefüttert. Hier wurden zusätzlich vor Eingabe in das Modell noch die Trainings und Testdaten mit einem MinMaxScaler skaliert, um eine mögliche höhere Genauigkeit durch Skalieren von Ausreißern zu bekommen. Der MinMaxScaler wurde gewählt, da er im Vergleich zu anderen Scalern die besten Werte produzierte. Hier kann man feststellen, dass die Genauigkeit der Trainingsdaten 0.58 und der Testdaten 0.55 daraufhin deuten, dass eine gröbere Unterteilung in Testklassen festgelegt werden muss. In diesem, und den folgenden Modellen, wurde nur die Genauigkeit betrachtet, da eine geringe Falsch-Negativ-Rate oder geringe Falsch-Positiv-Rate nicht ausschlaggebend ist und somit auch Recall und Precision nicht relevant sind.

Ein weiter Versuch in die bereits vorhandenen Qualitätsstufen zu unterteilen, wurde jedoch noch mit dem KNeighborsClassifier Modell unternommen. Es findet wieder derselbe Ablauf wie bei der vorherigen Methode statt. Dieses Modell gibt jedoch signifikant verbesserte Ergebnisse für die Trainingsdaten aus. Die Genauigkeit beträgt hier 0.72 für die Trainingsdaten und 0.58 für die Testdaten.

Um eine weitere Verbesserung zu erreichen, unterteilen wir nun die Qualität in 3 verschiedene Stufen: Schlecht, Mittel, Gut. Schlecht beinhaltet die Werte von 0 – 5, Gut beinhaltet 6 - 7 und Sehr Gut 8 – 10. Der benutzte Dataframe wird nun anhand dieser Qualitätsstufen aktualisiert, in Trainings- und Testdaten unterteilt, und mit einem MinMaxScaler skaliert.

Das Modell des KNeighborsClassifier wird nun initialisiert und die Trainingsdaten werden übergeben. Dieses ordnet nun den Daten die jeweils prognostizierte Qualitätsstufe (1-3) zu. Hier kommt man nun auf viel bessere Ergebnisse wie in Abbildung 10 zu sehen. Vor allem die Testdaten machten einen signifikanten Sprung und sind nun nutzbar.



Abbildung 10: Ergebnisse des KNeighbor-Verfahrens bei Einteilung der Qualität in drei Klassen

Im Anschluss darauf wird dieselbe Analyse mit 2 Klassen durchgeführt: Die untere Hälfte beinhaltet die Qualitäten 0 – 6, die obere Hälfte 7 - 10. Hier kommt man nun zu sehr guten Ergebnissen mit einer Genauigkeit für die Trainingsdaten von 0,91 und für die Testdaten von 0,89



Abbildung 11: Ergebnisse des KNeighbor-Verfahrens bei Einteilung der Qualität in zwei Klassen

Ein Grund warum diese Einteilungen so gut funktionieren könnte sein, dass eine große Anzahl der Weine eine Qualität von 0 bis 7 besitzt.

Diese beiden Einteilungen wurden auch mit Hilfe eines neuronalen Netzes untersucht. Die letzte Schicht der Neuronen besteht hierbei aus der Anzahl der Klassen und der Aktivierungsfunktion softmax. Dieses Modell wird dann mit dem loss als categorical\_crossentropy und der Genauigkeit als Metrik kompiliert. Bei Evaluierung der beiden Methoden stellt man fest, dass die Klassifikation mit 3 Klassen eine Genauigkeit der Trainingsdaten von 0,93 und der Testdaten von 0,88 zurückgibt.

Die Klassifikation mit 2 Klassen gibt für die Trainingsdaten eine Genauigkeit von 0,78 und für die Testdaten eine Genauigkeit von 0,76 zurück. Allerdings wurde hier eine batch\_size von 16 und eine Epochenanzahl von 200 verwendet.

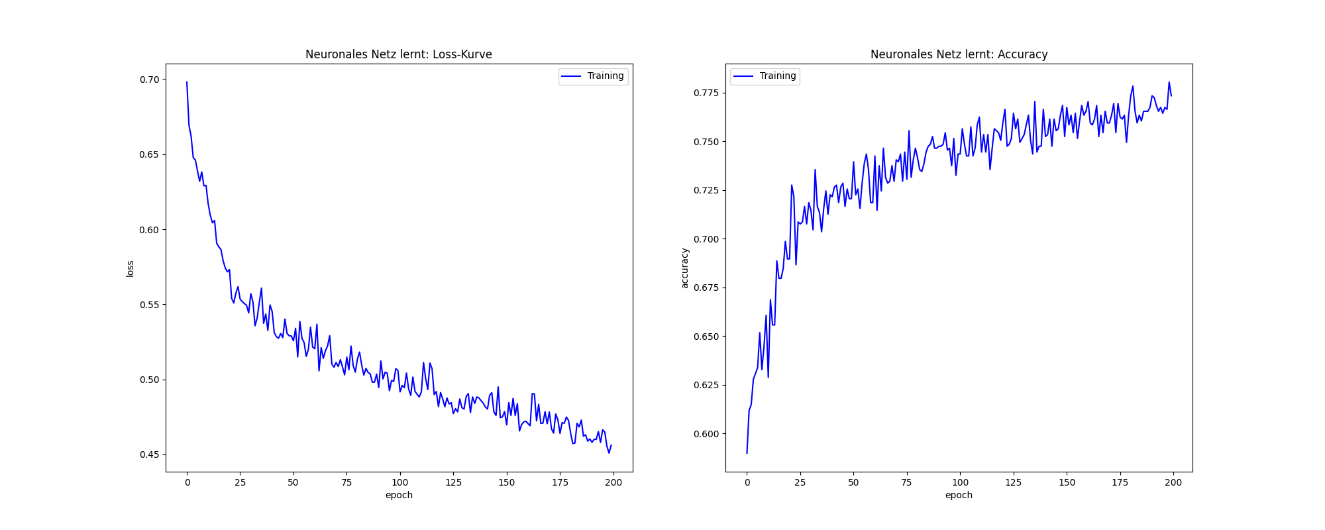


Abbildung 12: Lernkurve des neuronalen Netzes zur Klassifizierung nach 200 Epochen

Wenn man sich jedoch Abbildung 12 ansieht, erkennt man, dass die beiden Kurven des Loss und der Genauigkeit noch nicht vollständig konvergiert sind. Wenn man dasselbe Prozedere nun mit einer batch\_size von 32 und einer höheren Epochenanzahl anwendet sieht man, dass die Ergebnisse immer besser werden.

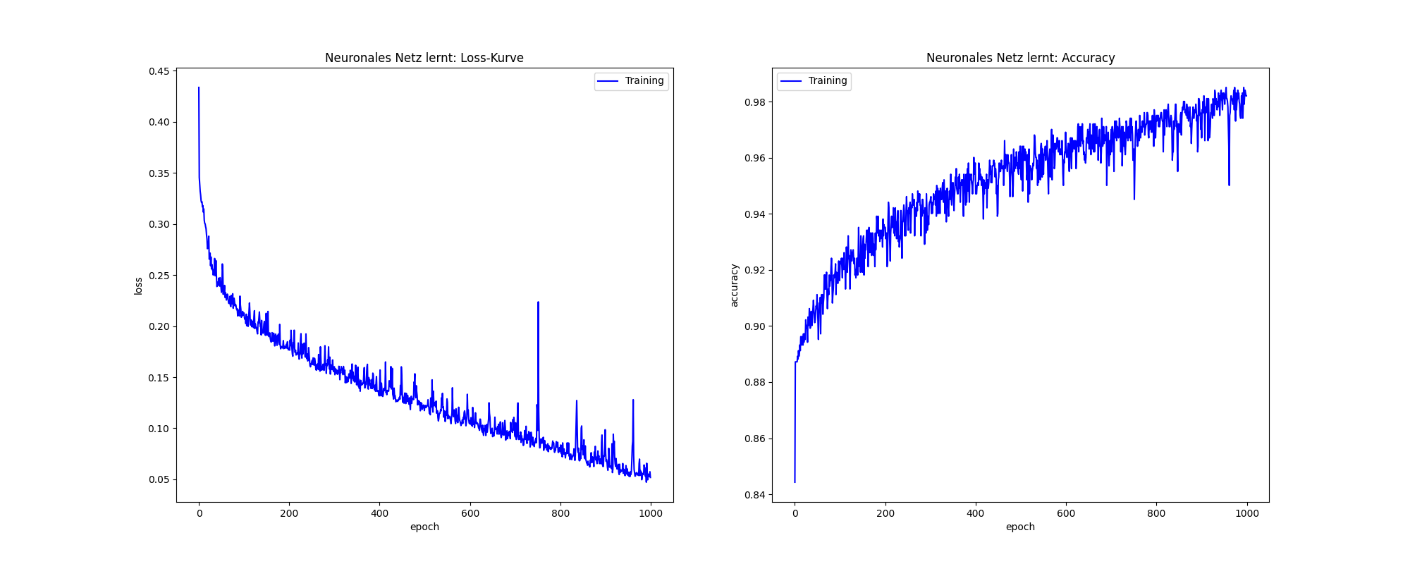


Abbildung 13: Lernkurve des neuronalen Netzes zur Klassifizierung nach 1000 Epochen

Wie man in Abbildung 13 sehen kann, ist eine Epochenanzahl von 1000 nahezu ideal. Dies ergibt eine Genauigkeit der Trainingsdaten von 0,97 und der Testdaten von 0,88. Für die Unterteilung in 2 Klassen resultiert dies in eine Genauigkeit der Trainingsdaten von 0,83 und der Testdaten von 0,75 für die Unterteilung in 3 Klassen. Ab einer höheren Epochenzahl beginnt ein Overfitting und die Genauigkeit für die Testdaten nimmt ab.

Die Klassifikation mit Entscheidungsbaum lieferte ebenfalls gute Ergebnisse, konnte die Ergebnisse der neuronalen Netze und des KNeighbour-Verfahren nicht übertreffen. Die genauen Ergebnisse finden sich in Abbildung 14.

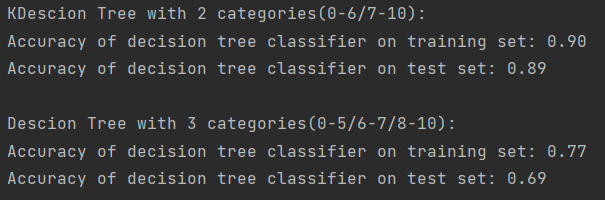


Abbildung 14: Ergebnisse der Klassifikation mit Entscheidungsbaum

## Erkenntnisse

Wie man anhand der vorhergehenden Daten feststellen kann, liefert die Klassifikation mithilfe von 2 bzw. 3 Klassen sehr gute Ergebnisse. Für Privatpersonen ist diese Klassifikation eher weniger sinnvoll, jedoch kann dies für einen Önologen oder Caterer zur Bewertung bzw. Einteilung von Weinen in Qualitätsklassen hilfreich sein.

# Segmentierung

Das Ziel der Segmentierung war es festzustellen, ob ein Modell den zusammengefügten Datensatz richtig in die Klassen rot und weiß einteilen kann. Als Modell wurde Kmeans mit 2 Clustern und einem Zufallsstrom von 12 ausgewählt. Wenn dieser Zufallsstrom nicht gesetzt wird, wird automatisch ein zufälliger ausgewählt. Dieser zufällige kann jedoch so ungünstig sein, dass ein um ein Vielfaches schlechteres Resultat herauskommt. Mit den gewählten Parametern kommt man auf eine insgesamte Genauigkeit von 0,77, einer Genauigkeit den Rotwein zu erkennen von 0,98 und einer Genauigkeit den Weißwein zu erkennen von 0,71. Eine mögliche Erklärung für die Einteilung in Rot- und Weißwein durch das Modell bei der Vorgabe von zwei Klassen liefert der unterschiedliche Säuregehalt von Rot- und Weißwein. Mit anderen Segmentierungsverfahren, wie hierarchischem Clustering, wurde keine Einteilung in Rot- und Weißweine erreicht.

# Assoziationsanalyse

Das Ziel der Assoziationsanalyse ist es Zusammenhänge zwischen den verschiedenen Spalten zu finden. Hierzu wurden die in der Vorlesung vorgestellten Verfahren verwendet. Da sich die Ergebnisse der einzelnen Verfahren überlappen, werden im Folgenden die gefundenen Zusammenhänge dargestellt. Jedem gefunden Zusammenhang werden unterstützende Ergebnisse der verschiedenen Verfahren beigelegt.

Die Assoziationsanalyse wurde nur für die Proben der Rotweine durchgeführt, da der Datensatz der Weißweine für vorgesehenen Umfang des Projekts zu groß ist. Es kann jedoch mindestens bei den chemischen Zusammenhängen ausgegangen werden, dass diese auch für die Weißweine gelten.

## Verwendete Verfahren

Um proportionale oder indirekt proportionale Zusammenhänge zu finden, wurden Scatter-Plots aus dem Paket pandas.plotting verwendet.

Um den Support, die Konfidenz oder Lift zweier Ereignisse zu berechnen werden die kontinuierlichen Daten diskretisiert. Dazu wurden die Werte jeder Spalte in eine Gruppenzugehörigkeit transformiert. Jedem kontinuierlichem Wert x eines Attributs wird dazu einer der folgenden Gruppen zugewiesen:

* Attribut\_low falls x <= q0.75(Attribut)
* Attribut\_mid falls q0.25(Attribut) < x <= q0.75(Attribut)
* Attribut\_high falls x > q0.75(Attribut)

qp(a) bezeichnet im hierbei das p-Quantil des Attributs a.

Zur Berechnung von Support, Konfidenz und Lift der verschiedenen Gruppen wurden die Methoden apriori und association\_rules aus dem Paket mlxtend.frequent\_patterns verwendet. Dabei wurden stets nur Ereignis-Paarung gewählt, die einen Support von mindestens 5% besitzten. D.h. es werden nur Ereignis-Paarungen ausgegeben, die bei mindestens 5% der ~1600 Proben auftreten. Somit ist sichergestellt, dass es sich bei den Ereignis-Paarungen nicht um Einzelfälle handelt. Anschließend wurden die Ereignis-Paarungen sortiert um solche mit möglichst hoher Konfidenz, mit möglichst hohem List, mit Lift nahe eins und möglichst kleinem Lift zu finden.

## Chemische Zusammenhänge

### Säureanteile - pH-Wert

In Abbildung 15 lässt sich klar eine Abhängigkeit des pH-Wertes von dem festen Säuregehalt und dem Citronensäure-Gehalt erkennen. Dies lässt sich einfach durch die Tatsache erklären, dass die Citronensäure eine feste Säure ist und somit Teil des festen Säuregehalts ist. Außerdem ist der pH-Wert per Definition kleiner je mehr Säure enthält.

Der Gehalt der flüchtigen Säureanteile verringert sich, wenn sich der Gehalt der festen Säureanteile bzw. der Citronensäure erhöht. Dies wird vermutlich absichtlich durch den Kelterer herbeigeführt, um einen zu säuerlichen Geschmack des Weines zu verhindern.

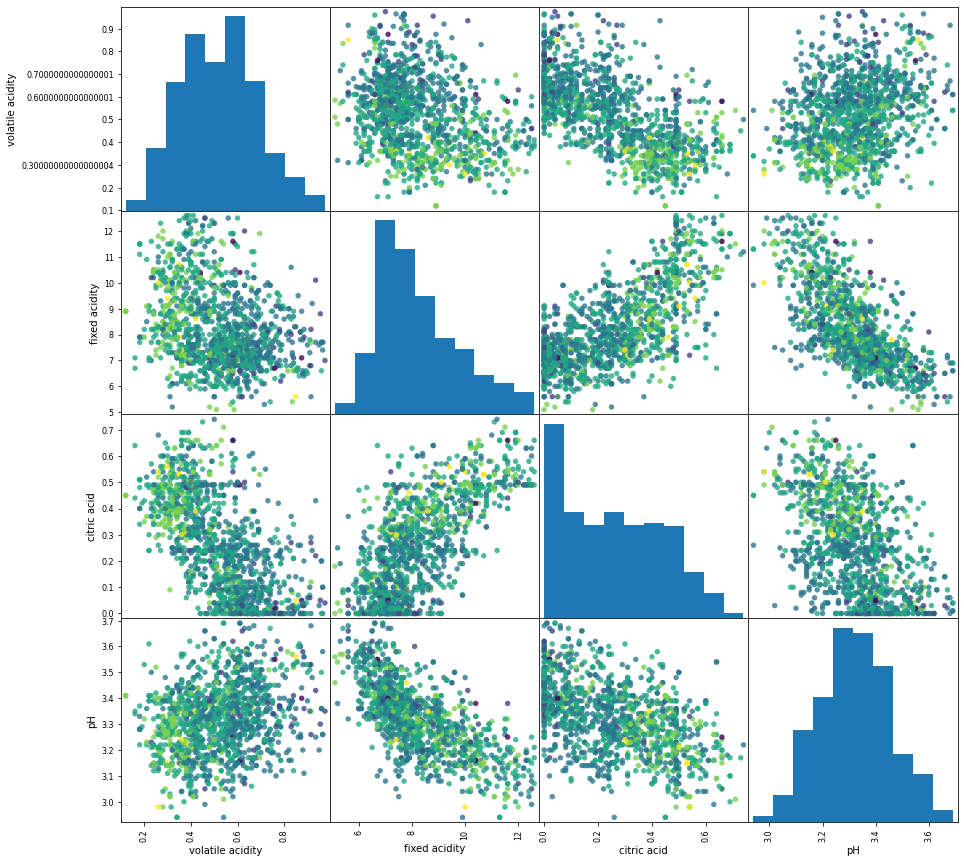


Abbildung 15: Scatter-Plot zum Säuregehalt und ph-Wert

Bestätigt werden die Zusammenhänge durch folgende Ergebnisse in Tabelle 1. Zur Verständlichkeit wird die erste Zeile beispielhaft erläutert. Der Support von 0.16 gibt an, dass jeder sechste Rotwein einen hohen festen Säureanteil sowie einen hohen Citronensäure-Anteil hat. Dies ist somit kein Einzelfall und erhöht die Verlässlichkeit der Aussage. Die Konfidenz von 68% gibt an, dass in rund ein 2/3 aller Fällen, in denen der feste Säuregehalt hoch ist, auch der Gehalt der Citronensäuren hoch ist. Der Lift von 2,74 gibt an, dass es knapp dreimal so wahrscheinlicher ist, dass der Gehalt der Citronensäuren hoch ist, falls ein Wein einen hohen Gehalt an festen Säuren besitzt.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X | Y | Support(X->Y) | Konfidenz(X->Y) | Lift(X->Y) |
| fixed acidity\_high | citric acid\_high | 0.166043 | 0.685185 | 2.742792 |
| citric acid\_high | fixed acidity\_high | 0.166043 | 0.664671 | 2.742792 |
| pH\_high | fixed acidity\_low | 0.160060 | 0.642643 | 2.447901 |
| pH\_low | fixed acidity\_high | 0.151832 | 0.598820 | 2.471057 |
| fixed acidity\_low | pH\_high | 0.160060 | 0.609687 | 2.447901 |
| fixed acidity\_high | density\_high | 0.148093 | 0.611111 | 2.446274 |
| fixed acidity\_mid | density\_mid | 0.294690 | 0.595166 | 1.205662 |
| citric acid\_high | volatile acidity\_low | 0.144353 | 0.577844 | 2.259000 |
| volatile acidity\_low | citric acid\_high | 0.144353 | 0.564327 | 2.259000 |
| citric acid\_low | volatile acidity\_high | 0.127898 | 0.510448 | 2.061839 |

Tabelle 1: Analyse des Säuregehalts

### Säureanteile - Alkohol – Dichte

Abbildung 16 kann man entnehmen das die Dichte mit steigendem dem Gehalt von festen Säuren zunimmt und mit steigendem Alkoholgehalt fällt. Dass sich die Dichte erhöht, je höher der feste Säureanteil ist, lässt darauf schließen, dass die festen Säurebestandteile schwerer sind der Durchschnitt der restlichen Inhaltsstoffe. Dass die Dichte geringer ist, je höher der Alkoholgehalt ist, lässt darauf schließen, dass der Alkohol leichter ist als der Durchschnitt der restlichen Bestandteile. Tabelle 2 bestätigt diese Ergebnisse.

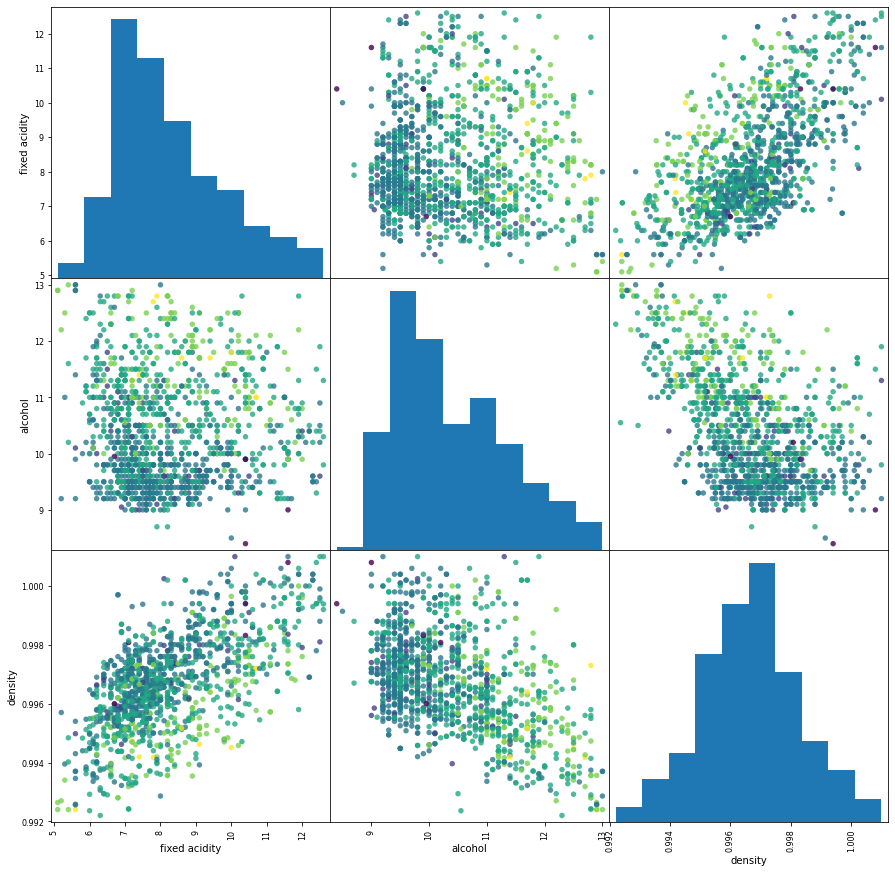


Abbildung 16: Scatter-Plot zu Weindichte

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X | Y | Support(X->Y) | Konfidenz(X->Y) | Lift(X->Y) |
| alcohol\_high | density\_low | 0.154076 | 0.637771 | 2.486005 |
| density\_low | alcohol\_high | 0.154076 | 0.600583 | 2.486005 |
| fixed acidity\_high | density\_high | 0.148093 | 0.611111 | 2.446274 |
| density\_high | fixed acidity\_high | 0.148093 | 0.592814 | 2.446274 |

Tabelle 2: Analyse der Weindichte

### Schwefeldioxide

In Abbildung 17 lässt sich zweifelfrei ein Zusammenhang zwischen dem Gehalt freier Schwefeldioxide und dem Gesamtgehalt an Schwefeldioxide erkennen. Da der Anteil der freien Schwefeldioxide zu dem Gesamtgehalt beiträgt ist dies nur logisch. Der Zusammenhang hat sich in der Berechnung des Supportes, der Konfidenz und des Lifts bestätigt. Die Ergebnisse der Berechnungen findet sich in Tabelle 3.

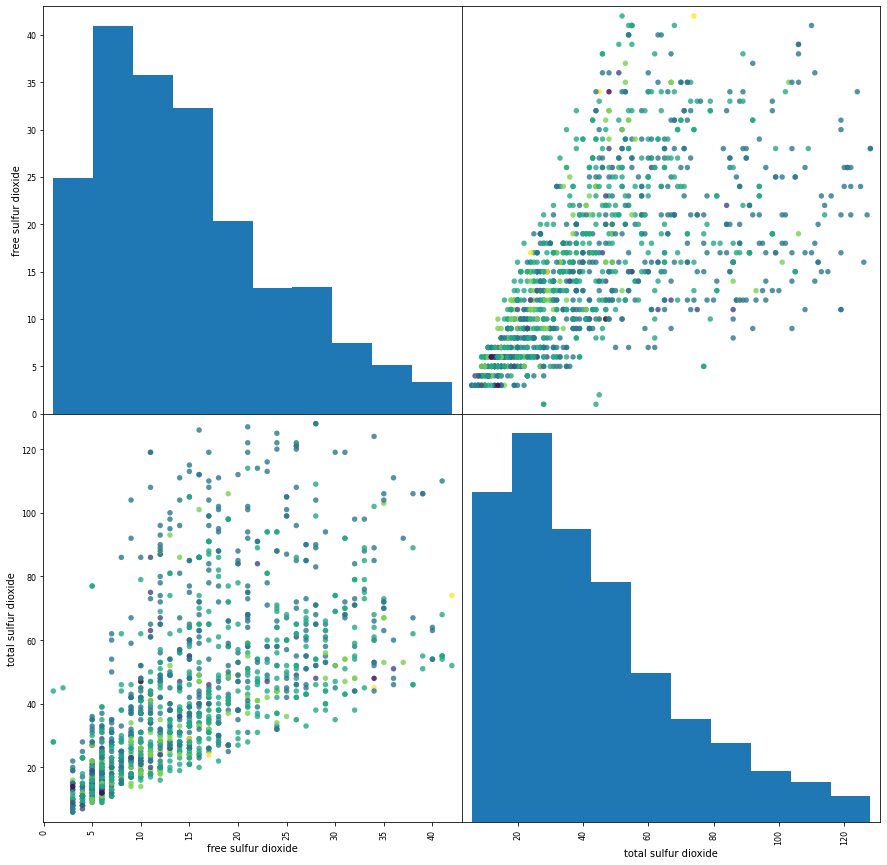


Abbildung 17: Scatterplot zum Schwefelgehalt

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X | Y | Support(X->Y) | Konfidenz(X->Y) | Lift(X->Y) |
| total sulfur dioxide\_low | free sulfur dioxide\_low | 0.192221 | 0.758112 | 2.963731 |
| free sulfur dioxide\_low | total sulfur dioxide\_low | 0.192221 | 0.751462 | 2.963731 |

Tabelle 3: Analyse des Schwefelgehalts

## Qualitative Zusammenhänge

Qualitätsmindernde Eigenschaften

Aus Tabelle 4 kann man folgern, dass zumindest die folgenden Eigenschaften oft mit einer geringeren Qualität einhergehen und daher bei der Wahl eines vindo verde zu vermeiden sind:

* Geringer Sulfatgehalt
* Geringer Alkoholgehalt
* Geringer Gehalt flüchtiger Säuren
* Hoher Schwefeldioxidgehalt

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X | Y | Support(X->Y) | Konfidenz(X->Y) | Lift(X->Y) |
| sulphates\_low | quality\_low | 0.183994 | 0.702857 | 1.532985 |
| alcohol\_low | quality\_low | 0.185490 | 0.714697 | 1.558810 |
| volatile acidity\_high | quality\_low | 0.158564 | 0.640483 | 1.396943 |
| total sulfur dioxide\_high | quality\_low | 0.157816 | 0.635542 | 1.386166 |

Tabelle 4: Qualitätsmindernde Eigenschaften

Die Eigenschaften aus Tabelle 5 haben alle einen Lift nahe bei eins. D.h. die folgenden Eigenschaften sind keine Indikatoren für eine verminderte Qualität:

* Niedriger pH-Wert
* Hoher oder niedriger Restzucker
* Hoher Gehalt freier Schwefeldioxide

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X | Y | Support(X->Y) | Konfidenz(X->Y) | Lift(X->Y) |
| pH\_low | quality\_low | 0.113687 | 0.448378 | 0.977946 |
| residual sugar\_high | quality\_low | 0.115931 | 0.476923 | 1.040206 |
| residual sugar\_low | quality\_low | 0.142857 | 0.478697 | 1.044074 |
| free sulfur dioxide\_high | quality\_low | 0.118175 | 0.478788 | 1.044273 |

Tabelle 5: Eigenschaften minder qualitativer Weine

### Qualitätsfördernde Eigenschaften

Die Eigenschaften aus Tabelle 6 sind bedingt Indikatoren für eine hohe Weinqualität. Sie besitzen eine niedrige Konfidenz, d.h. konkret für die erste Zeile, nur ein Drittel der Weine mit hohem Alkoholgehalt sind auch hoch qualitative Weine. Jedoch ist die Wahrscheinlichkeit, dass ein Wein eine hohe Qualität besitzt, doppelt so hoch, falls er einen hohen Alkoholgehalt hat.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X | Y | Support(X->Y) | Konfidenz(X->Y) | Lift(X->Y) |
| alcohol\_high | quality\_high | 0.081526 | 0.337461 | 2.654034 |
| volatile acidity\_low | quality\_high | 0.074046 | 0.289474 | 2.276625 |
| sulphates\_high | quality\_high | 0.071055 | 0.288754 | 2.270964 |
| density\_low | quality\_high | 0.062827 | 0.244898 | 1.926050 |

Tabelle 6: Eigenschaften hoch qualitativer Weine

Die Eigenschaften aus Tabelle 7 sind Indikatoren dafür, dass ein Wein nicht minderer Qualität hat. D.h. wählt man einen Wein mit den Eigenschaften aus Tabelle 7 besteht eine höhere Wahrscheinlichkeit, es nicht um einen Wein minderer Qualität handelt. Da der Support und die Konfidenzen niedrig sind, sind die Aussagen auf weniger Weinproben als die vorherigen Ergebnisse gestützt.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X | Y | Support(X->Y) | Konfidenz(X->Y) | Lift(X->Y) |
| sulphates\_high | quality\_low | 0.053852 | 0.218845 | 0.477318 |
| volatile acidity\_low | quality\_low | 0.061331 | 0.239766 | 0.522948 |
| density\_low | quality\_low | 0.069559 | 0.271137 | 0.591371 |

Tabelle 7: Fehlende Eigenschaften minder qualtiativer Weine

## Erkenntnisse

Die Assoziationsanalyse hat verschiedene chemische Zusammenhänge der Attribute aufgezeigt. Eine genaue Erklärung für diese erfordert jedoch das Wissen eines Fachmanns. Weiter wurden Indikatoren gefunden die auf einen Rotwein minderer Qualität hinweisen. Es existieren auch Indikatoren, die auf eine gehobene Weinqualität hinweisen. Letztere werden aber nur von einer geringen Anzahl an Weinproben gestützt. Die meisten Indikatoren sind nicht als Unterstützung des Konsumenten beim Weinkauf geeignet, dass die benötigten Kennzahlen dem Konsumenten nicht vorliegen. Sie können jedoch eine Hilfestellung für Önologen bei der Bewertung von Weinen bieten.

# Fazit

An dem vorliegenden Datensatz wurden alle Schritte einer Datenanalyse mithilfe verschiedener Techniken durchgeführt. Eine unerwartete Hürde zu Beginn war die Beschaffung von brauchbaren Daten, die aus einer verlässlichen Quelle stammen. Der ausgewählte Datensatz beinhaltet verschiedene Weinproben für welche mehrere chemische Kennzahlen sowie eine Qualitätsbewertung vorliegen. Eine Schätzung der Qualität durch Regressionsverfahren schlug fehl. Durch die Diskretisierung der Qualitätsstufen in größere Gruppen und die Nutzung von Klassifikationsverfahren zur Schätzung der Qualität konnte eine zuverlässiges Prognosemodell erstellt werden. Mithilfe der Assoziationsanalyse konnten verschiedene chemische Zusammenhänge und Indikatoren für geringere bzw. höhere Weinqualität aufgezeigt werden.

# Abbildungsverzeichnis

[Abbildung 1: Box-Plot aller Attribute der Rotweine 4](#_Toc92873389)

[Abbildung 2:Box-Plot aller Attribute der Weißweine 5](#_Toc92873390)

[Abbildung 3: Scattermatrix Rotwein 6](#_Toc92873391)

[Abbildung 4: Ergebnisse der linearen Regression bei Rotweinen 7](#_Toc92873392)

[Abbildung 5: Ergebnisse der polynomiellen Regression bei Rotwein 7](#_Toc92873393)

[Abbildung 6: Ergebnisse der Support-Vector-Regression bei Rotweinen 8](#_Toc92873394)

[Abbildung 7: Ergebnisse der Entscheidungsbaum-Regression bei Rotweinen 9](#_Toc92873395)

[Abbildung 8: Ergebnisse der Regression mit neuronalen Netzen bei Rotweinen 9](#_Toc92873396)

[Abbildung 9: Lernkurve der Neuronalen Netze bei Regression von Rotweinen 10](#_Toc92873397)

[Abbildung 10: Ergebnisse des KNeighbor-Verfahrens bei Einteilung der Qualität in drei Klassen 11](#_Toc92873398)

[Abbildung 11: Ergebnisse des KNeighbor-Verfahrens bei Einteilung der Qualität in zwei Klassen 11](#_Toc92873399)

[Abbildung 12: Lernkurve des neuronalen Netzes zur Klassifizierung nach 200 Epochen 12](#_Toc92873400)

[Abbildung 13: Lernkurve des neuronalen Netzes zur Klassifizierung nach 1000 Epochen 12](#_Toc92873401)

[Abbildung 14: Ergebnisse der Klassifikation mit Entscheidungsbaum 13](#_Toc92873402)

[Abbildung 15: Scatter-Plot zum Säuregehalt und ph-Wert 15](#_Toc92873403)

[Abbildung 16: Scatter-Plot zu Weindichte 16](#_Toc92873404)

[Abbildung 17: Scatterplot zum Schwefelgehalt 17](#_Toc92873405)

# Tabellenverzeichnis

[Tabelle 1: Analyse des Säuregehalts 16](#_Toc92873406)

[Tabelle 2: Analyse der Weindichte 17](#_Toc92873407)

[Tabelle 3: Analyse des Schwefelgehalts 17](#_Toc92873408)

[Tabelle 4: Qualitätsmindernde Eigenschaften 18](#_Toc92873409)

[Tabelle 5: Eigenschaften minder qualitativer Weine 18](#_Toc92873410)

[Tabelle 6: Eigenschaften hoch qualitativer Weine 18](#_Toc92873411)

[Tabelle 7: Fehlende Eigenschaften minder qualtiativer Weine 19](#_Toc92873412)

# Literaturverzeichnis

|  |  |
| --- | --- |
| [1] | OIV, „Leading countries in wine export worldwide in 2020, based on volume (in million hectoliters),“ 17 Mai 2021. [Online]. Available: https://www.statista.com/statistics/240649/top-wine-exporting-countries-since-2007/. [Zugriff am 11 Januar 2022]. |
| [2] | A. C. F. A. T. M. a. J. R. P. Cortez, „Modeling wine preferences by data mining from physicochemical properties,“ *Descion Support Systems,* Bde. %1 von %2 47(4):547-553, Nr. 2009, p. Descion Support Systems, P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.. |