**Projektbericht zum Modul Information Retrival und Visualisierung**

Sommersemester 2021

**Visualisierung von verschiedenen Weindaten**

Marie-Luise Lindner

Matrikelnummer: 211231747

GitHub Repository: <https://github.com/RicBre/Elm-Projekt-WineInformation>

# Inhaltsverzeichnis

[Inhaltsverzeichnis II](#_Toc82161120)

[1. Einleitung 1](#_Toc82161121)

[1.1 Anwendungshintergrund 1](#_Toc82161122)

[1.2 Zielgruppen 3](#_Toc82161123)

[1.3 Überblick und Beiträge 4](#_Toc82161124)

[2. Daten 5](#_Toc82161125)

[2.1 Technische Breitstellung der Daten 6](#_Toc82161126)

[2.2 Datenvorverarbeitung 6](#_Toc82161127)

[3. Visualisierung 7](#_Toc82161128)

[3.1 Analyse der Anwendungsaufgaben 7](#_Toc82161129)

[3.2 Anforderungen an die Visualisierungen 8](#_Toc82161130)

[3.3 Präsentation der Visualisierung 9](#_Toc82161131)

[3.3.1 Visualisierung Eins 9](#_Toc82161132)

[3.3.2 Visualisierung Zwei 10](#_Toc82161133)

[3.3.3 Visualisierung Drei 12](#_Toc82161134)

[3.4 Interaktion 12](#_Toc82161135)

[4. Implementierung 13](#_Toc82161136)

[5. Anwendungsfälle 14](#_Toc82161137)

[5.1 Anwendung Visualisierung Eins 14](#_Toc82161138)

[5.2 Anwendung Visualisierung Zwei 14](#_Toc82161139)

[5.3 Anwendung Visualisierung Drei 16](#_Toc82161140)

[6. Verwandte Arbeiten 17](#_Toc82161141)

[7. Zusammenfassung und Ausblick 18](#_Toc82161142)

[8. Literatur i](#_Toc82161143)

[Anhang ii](#_Toc82161144)

[Git Historie ii](#_Toc82161145)

# Einleitung

Bereits zur Zeit des antiken Griechenlands herrschte der Wunsch nach einer Maschine mit der Fähigkeit menschlichen Denkens. Mehr als einhundert Jahre bevor programmierbare Computer wissenschaftliche Wirksamkeit erzielten, stellte sich die Frage nach künstlicher Intelligenz (KI)[[1]](#footnote-2). Heute ist künstliche Intelligenz fester Bestandteil des menschlichen Alltags. Es ist ein blühendes Feld mit vielen praktischen Anwendungen und aktiven Forschungsthemen. Im bestehenden gesellschaftlichen Kontext bildet intelligente Software die Basis für automatisierte Routinearbeiten, das Verständnis von Sprache oder Bildern, Diagnosen in der Medizin, sowie der Unterstützung wissenschaftlicher Grundlagenforschung.[[2]](#footnote-3)

Automatisierte Routinearbeiten sind unter anderem im Sektor des Automobilvertriebes beobachtbar. Berichten prognostizieren ein Wachstum bei Neuwagen in den kommenden 5 Jahren durchschnittlich um 3,5 %. Ein Anstieg im Gebrauchtwagensektor sei von 5% zu registrieren. Kausal der gesteigerten Nachfrage des Gebrauchtwagenmarktes sind veränderte Bedürfnisstrukturen der Verbraucher zu verzeichnen. Primäres Ziel ist die Optimierung des Wiederverkaufspreises das Automobiles. Die Validierung dessen ist verbunden mit Werkstattbesuchen, sowie Kostenvoranschlägen und wird als zeit -und ressourcenaufwändiger Prozess betrachtet. Ein indisches Unternehmen namens „avl“ fokussiert diese Thematik in einem Projekt. Ziel ist die Entwicklung eines Modells für Drittunternehmen, zur Schätzung des Preises des Autos des Kunden im Online-Portal. Folglich erfolge eine Rationalisierung des Verkaufsprozesses von Gebrauchtwagen. Methodisch soll die Umsetzung dieses Projektes mit Hilfe von maschinellem Lernen auf Basis eines Trainingsdatensatzes von rund 7000 Daten über Gebrauchtwagen erfolgen.[[3]](#footnote-4)

Der vorliegende Bericht nimmt Rückschlüsse auf diesen Sachverhalt und die vorliegenden Daten. Ziel ist die Aufarbeitung des Datensatzes, sowie dessen anschließende Visualisierung in drei differenzierten Visualisierungstechniken.

## Anwendungshintergrund

Um ein Verständnis für die Wichtigkeit einer signifikanten Datengrundlage für künstliche Intelligenzen zu schaffen, steht im Folgenden der Anwendungshintergrund im Fokus.

Die Anfänge der Entwicklung von KI’s fokussierten die Umsetzung von Problemen, die sich durch eine Liste formaler, mathematischer Regeln beschreiben lassen. Als größte Herausforderung für die künstliche Intelligenz stellten sich Fragen heraus, welche der Mensch intuitiv und automatisiert löst, wie das Erkennen von Gesichtern in Bildern. Ansätze zur Lösung sind künstliche neuronale Netze oder Entscheidungsbäume. Diese sind von der Funktionsweise des menschlichen Gehirns inspiriert. Sie ermöglichen Computern aus Erfahrungen zu lernen, indem sie die Welt in Form einer Hierarchie von Konzepten verstehen. Jedes Konzeptebene ist durch die Beziehung zu einfacheren Konzepten definiert. Die komplexeren Hierarchieebenen lernen von den einfacheren. Basierend dieses Grundsatzes ist die Wissensaneignung der KI-Systeme durch das Extrahieren von Mustern aus Rohdaten, sogenannten Trainingsdaten, möglich. Diese Fähigkeit ist als maschinelles Lernen

definiert. Mittels dessen können Computer Sachverhalte lösen, die Kenntnisse über die reale Welt erfordern oder Entscheidungen treffen, die subjektiv erscheinen.

Der Sektor der künstlichen Intelligenz wird unterschieden in deskriminative und generative Modelle. Deskriminative Modelle sind darauf reduziert bedingte Wahrscheinlichkeiten zu lernen.[[4]](#footnote-5) Generative Modelle hingegen lernen nicht die bedingte Wahrscheinlichkeit, sondern fokussieren die gesamte Wahrscheinlichkeitsverteilung. Dies befähigt das System der Zuweisung von Eigenschaften zu Objekten und folglich deren Erzeugung. Entsprechend findet dieses Modell der Datenverarbeitung vermehrt Anwendung in der Automatisierung und ist die Methodik im vorliegenden Beispiel der Erstellung eines Online-Portals zur Rationalisierung des Verkaufsprozesses von Gebrauchtwagen.

Schlussfolgernd lernt eine intelligente Software ausschließlich über die Einspeisung von Daten. Analog spiegeln die hinterlegten Trainingsdaten das Level an Intelligenz des Systems wider. Ein Bias des Datensatzes führt simultan zu einer Verzerrung der KI. Infolgedessen ist das Aufarbeiten des Trainingsdatensatzes höchste Priorität. Aus diesem Grund erfolgt im vorliegenden Bericht eine genaue Analyse des Trainingsdatensatzes. Die Informationen innerhalb des Datensatzes werden aufgearbeitet, sodass eine Visualisierung dessen ermöglicht wird. Hauptziel ist es, dass unterschiedliche Zielgruppen, auf dieser Grundlage Zusammenhänge, Ausreißer, sowie signifikante Werte des Datensatzes erkennen und entsprechende Rückschlüsse auf das Etablieren der Trainingsdaten in das KI-System ziehen können. Darauf basierend wurden drei Visualisierungstechniken ausgewählt. Diese besitzen zu Teilen die Disposition der Interaktivität. Dies ermöglicht dem Nutzer den Vergleich von mehr als 2 Eigenschaften der gewählten Daten. Im Detail fanden die Visualisierungstechniken der Baumhierarchie, des Scatterplots und der Parallelen Koordinaten Anwendung. Diese werden nachfolgend kurz vorgestellt.

Bei der ersten Visualisierungstechnik handelt es sich um einen Scatterplot. Mit wird die Gegenüberstellung zweier verschiedener numerischer Variablen der Gebrauchtwagen ermöglicht. Dies bietet die Möglichkeit die Beziehung zwischen diesen Variablen zu beobachten. Korrelationsbeziehungen, Muster oder Trends können erkannt und analysiert werden. Additional können mit Hilfe dieser Technik Daten gruppiert werden, welches Rückschlüsse auf mögliche Ausreißer oder Datenlücken vereinfacht.[[5]](#footnote-6)

Die zweite Visualisierungstechnik repräsentiert den Ansatz der Parallelen Koordinaten. Primärer Vorteil dieses stilistischen Mittels ist die Möglichkeit der Präsentation mehrdimensionaler Daten in einem zweidimensionalen Raum.

## Zielgruppen

## Überblick und Beiträge

In diesem Projekt wurden die Daten von der Webseite „Kaggle“ verwendet. Dabei handelt es sich Informationen rund um das Thema Wein. Diese Daten enthalten dabei Auskünfte über den Namen, Produzent, Herkunft, Typ und Verwendung des Weines. Darüber hinaus gibt es Angaben zu den verschiedenen Merkmalen des Weines wie Alkoholgehalt, Trinktemperatur, Süße, Säure, Körper, Gerbstoffe, Preis, Jahr und größer einer Flasche. Genauere Informationen, was diese Eigenschaften ausdrücken, ist im Kapitel Das Hauptziel ist dabei, diesen Datensatz so bereitzustellen, dass die unterschiedlichen Zielgruppen ihre individuellen Fragen rund um das Thema Weine beantworten können. Um dies zu erfüllen, sollten zwei andersartige Charakteristika des Datensatzes gegenübergestellt werden können. Somit wäre es beispielsweise denkbar, das Preisleistungsverhältnis bei den unterschiedlichen Merkmalen herauszufinden. Weiterhin sollte es möglich sein, mehr als nur zwei Charakteristika Datensatzes zu vergleichen, um so ein besseres Verständnis für die Eigenschaften und Zusammenhänge des Weines zu schaffen. Außerdem sollten die Visualisierungen eine Möglichkeit der Zuordnung der Weine nach der Herkunft ermöglichen, umso beispielsweise überprüfen zu können, ob die verschiedenen Produktionsmengen mit den Weinarten eines Landes gleichmäßig mitwachsen. Diese Ziele werden anschließend in diesem Projektbericht mithilfe von drei Visualisierungstechniken bearbeitet.

Anwendungshintergrund zu finden. Diese Daten werden anschließend anhand ihrer verschiedenen Merkmale in einem Scatterplot, Parallelen Koordinaten und einer Baumhierarchie dargestellt. Wie diese Diagramme aufgebaut sind, wird im Kapitel Das Hauptziel ist dabei, diesen Datensatz so bereitzustellen, dass die unterschiedlichen Zielgruppen ihre individuellen Fragen rund um das Thema Weine beantworten können. Um dies zu erfüllen, sollten zwei andersartige Charakteristika des Datensatzes gegenübergestellt werden können. Somit wäre es beispielsweise denkbar, das Preisleistungsverhältnis bei den unterschiedlichen Merkmalen herauszufinden. Weiterhin sollte es möglich sein, mehr als nur zwei Charakteristika Datensatzes zu vergleichen, um so ein besseres Verständnis für die Eigenschaften und Zusammenhänge des Weines zu schaffen. Außerdem sollten die Visualisierungen eine Möglichkeit der Zuordnung der Weine nach der Herkunft ermöglichen, umso beispielsweise überprüfen zu können, ob die verschiedenen Produktionsmengen mit den Weinarten eines Landes gleichmäßig mitwachsen. Diese Ziele werden anschließend in diesem Projektbericht mithilfe von drei Visualisierungstechniken bearbeitet.

Anwendungshintergrund genauer erklärt. Dabei ist es möglich, mithilfe des Scatterplots zwei verschiedene Merkmale des Weines zu vergleichen, um entsprechende Muster zwischen diesen beiden Eigenschaften zu erkennen und identifizieren. Dies könnte zu neuen Erkenntnissen rund um diese Kriterien führen. Mithilfe der Parallelen Koordinaten können mehr als zwei Eigenschaften miteinander vergleichen werden. Dabei können verschiedenartige Trends und Besonderheiten bei den verschiedenen Merkmalen des Weines erkannt werden. Bei der Baumhierarchie wird es möglich sein, unterschiedliche Weine ihren Regionen zuzuordnen. Diese Informationen kann dann wiederum mit anderen Daten abgeglichen werden, um einen weiteren Erkenntnisgewinn zu ermöglichen.

# Daten

.

## Technische Breitstellung der Daten

.

## Datenvorverarbeitung

.

## Analyse der Anwendungsaufgaben

## Anforderungen an die Visualisierungen

## Präsentation der Visualisierung

### Visualisierung Eins

Abbildung : Scatterplot (Quelle: eigene Darstellung)

### Visualisierung Zwei

Abbildung 2: Parallele Koordinaten (Quelle: eigene Darstellung)

### Visualisierung Drei

Abbildung : Auszug aus der Baumhierarchie

## Interaktion

# Implementierung

# Anwendungsfälle

## Anwendung Visualisierung Eins

Abbildung 4: Scatterplot welcher die Eigenschaften Preis und Körper gegenüberstellt (Quelle: eigene Darstellung)

## Anwendung Visualisierung Zwei

Abbildung : Gegenüberstellung der Eigenschaften Körper, Gerbstoffe, Süße und Säuregehalt mithilfe der Parallelen Koordinaten (Quelle: eigene Darstellung)

## Anwendung Visualisierung Drei

Abbildung 6: Ausschnitt mit Asiatischen Weinen aus der Baumhierarchie (Quelle: eigene Darstellung)

# Verwandte Arbeiten

# Zusammenfassung und Ausblick

# Literaturverzeichnis

1. **Lovelace, A.** *Notes upon L. F. Menabrea’s “Sketch of the Analytical Engineinvented by Charles Babbage”.* 1842.

2. **Goodfellow, Ian, Bengio, Yoshua und Courville, Aaron.** *Deep Learning, Das umfassende Handbuch. Grundlagen, aktuelle Verfahren und Algorithmen, neue Forschungsansätze.* s.l. : MITP-Verlag, 2018. S. 912.

3. **https://www.kaggle.com/avikasliwal/used-cars-price-prediction?select=train-data.csv.** https://www.kaggle.com/avikasliwal/used-cars-price-prediction?select=train-data.csv. [Online] 26.11.2021. [Zitat vom: 26. 11 2021.] https://www.kaggle.com/avikasliwal/used-cars-price-prediction?select=train-data.csv.

# Anhang

## Git Historie

1. (1) [↑](#footnote-ref-2)
2. (2) [↑](#footnote-ref-3)
3. (3) [↑](#footnote-ref-4)
4. (4) [↑](#footnote-ref-5)
5. (5) [↑](#footnote-ref-6)