# Klassifizierung für E-Commerce Nutzergruppen auf Basis von Data Mining unter Verwendung von künstlicher Intelligenz.

## Untersuchung der Nutzerklassifizierung durch Datenanalyse




vorgelegte Masterarbeit
zum Erlangen des akademischen Grades
**Master of Science (M.Sc.)**

eingereicht von: Wilfried Pahl
Matrikelnummer: 901932
Studiengang: Online Medieninformatik
Berliner Hochschule für Technik

**Betreuender Prüfer** Prof. Dr. S. Edlich Berliner Hochschule für Technik
**Zweitgutachter** noch nicht bekannt auch von einer Hochschule

Temmen-Ringenwalde, der 4. Dezember 2022

## Stichworte

Data-Mining, Big Data, künstliche Intelligenz, Clusterung, Nutzergruppen, k-Means-Algorithm.

## Kurzzusammenfassung

Hier kommt das Abstract auf Deutsch.

## Keywords

Data-Mining, Big Data, künstliche Intelligenz, Clusterung, Nutzergruppen, k-Means-Algorithm.

## Abstract

Here comes later the abstract on english.

# Danksagung

Hier sage ich auch mal zu irgendjemand Danke.

# Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst habe, dass ich sie zuvor an keiner anderen Hochschule und in keinem anderen Studiengang als Prüfungsleistung eingereicht habe und dass ich keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Alle Stellen der Arbeit, die wörtlich oder sinngemäß aus Veröffentlichungen oder aus anderweitigen fremden Äußerungen entnommen wurden, sind als solche kenntlich gemacht.

Datum Unterschrift

# Inhaltsverzeichnis

**Inhaltsverzeichnis** viii

**Abbildungsverzeichnis** ix

**Tabellenverzeichnis** xi

**1 Einführung** 1
- 1.1 Motivation . . . 2
  - 1.1.1 Ausgangslage . . . 2
  - 1.1.2 Vorteile durch die Vorhersage . . . 2
- 1.2 Ziel der Arbeit . . . 2
- 1.3 Inhaltlicher Aufbau der Arbeit . . . 3

**2 Grundlagen** 5
- 2.1 Grundbegriffe . . . 5
- 2.2 Verwandte Arbeiten . . . 6
- 2.3 Big Data . . . 6
  - 2.3.1 Technologien . . . 8
- 2.4 Data-Mining . . . 8
- 2.5 Clustering . . . 8

**3 Explorative Datenanalyse** 9
- 3.1 Arten der EDA . . . 9
- 3.2 Methoden der Datenanalyse . . . 10
- 3.3 Kundengruppen . . . 10
  - 3.3.1 DiSG Modell . . . 10
  - 3.3.2 Mögliche Kundengruppen . . . 11
- 3.4 Daten für die Onlineshop-Nutzeranalyse . . . 12

**4 Maschinelles Lernen** 13
- 4.1 Methoden maschinellen Lernens . . . 13
  - 4.1.1 Überwachtes Lernen . . . 13
  - 4.1.2 Unüberwachtes Lernen . . . 13
  - 4.1.3 Teilweise überwachtes Lernen . . . 14
  - 4.1.4 Bestärktes Lernen . . . 14
- 4.2 Algorithmen zum Analysieren von Onlineshop Daten . . . 14
  - 4.2.1 Regressionsalgorithmen . . . 14
  - 4.2.2 Instanz-basierende Algorithmen . . . 15
  - 4.2.3 Regularisierungsalgorithmen . . . 15

4.2.4 Entscheidungsbaumalgorithmen . . . . . 15
4.2.5 Bayessche Algorithmen . . . . . 16
4.2.6 Clustering-Algorithmen . . . . . 16
4.2.7 Lernalgorithmen für Assoziationsregeln . . . . . 16
4.2.8 Neuronale Netze . . . . . 16
4.2.9 Depp-Learning-Algorithmen . . . . . 17
4.2.10 Dimensionsreduktionsalgorithmen . . . . . 17
4.2.11 Ensemble-Algorithmen . . . . . 17
4.3 Umsetzung der Datenverarbeitung . . . . . 18
4.4 Umsetzung der Clusterung . . . . . 18

**5 Evaluation** **19**
5.1 Aufbau der Umgebung . . . . . 19
5.2 Ergebnisse . . . . . 19
5.3 Bewertung und Diskussion . . . . . 19

**6 Future Work** **21**

**Anhang** **23**

**Literaturverzeichnis** **25**

**Glossar** **26**

# Abbildungsverzeichnis

1.1 Umsatzentwicklung E-Commerce von 2001-2022 . . . . . 1
2.1 Umsatzprognose E-Commerce bis 2025 . . . . . 7
3.1 DiSG Übersicht https://www.disg-modell.de . . . . . 10
4.1 Anzahl der Bestellung in Abhängigkeit der Warenkorbhöhe . . . . . 18

# Tabellenverzeichnis

# Kapitel 1

# Einführung

Das Volumen der bestellten Waren im E-Commerce Bereich wächst stetig und durch die Corona-Krise erfuhr dieser Prozess eine Beschleunigung. Die meisten Umsätze verzeichnen virtuelle Marktplätze, wie Amazon und Ebay, aber auch kleine Onlineshops können durchaus bestehen.

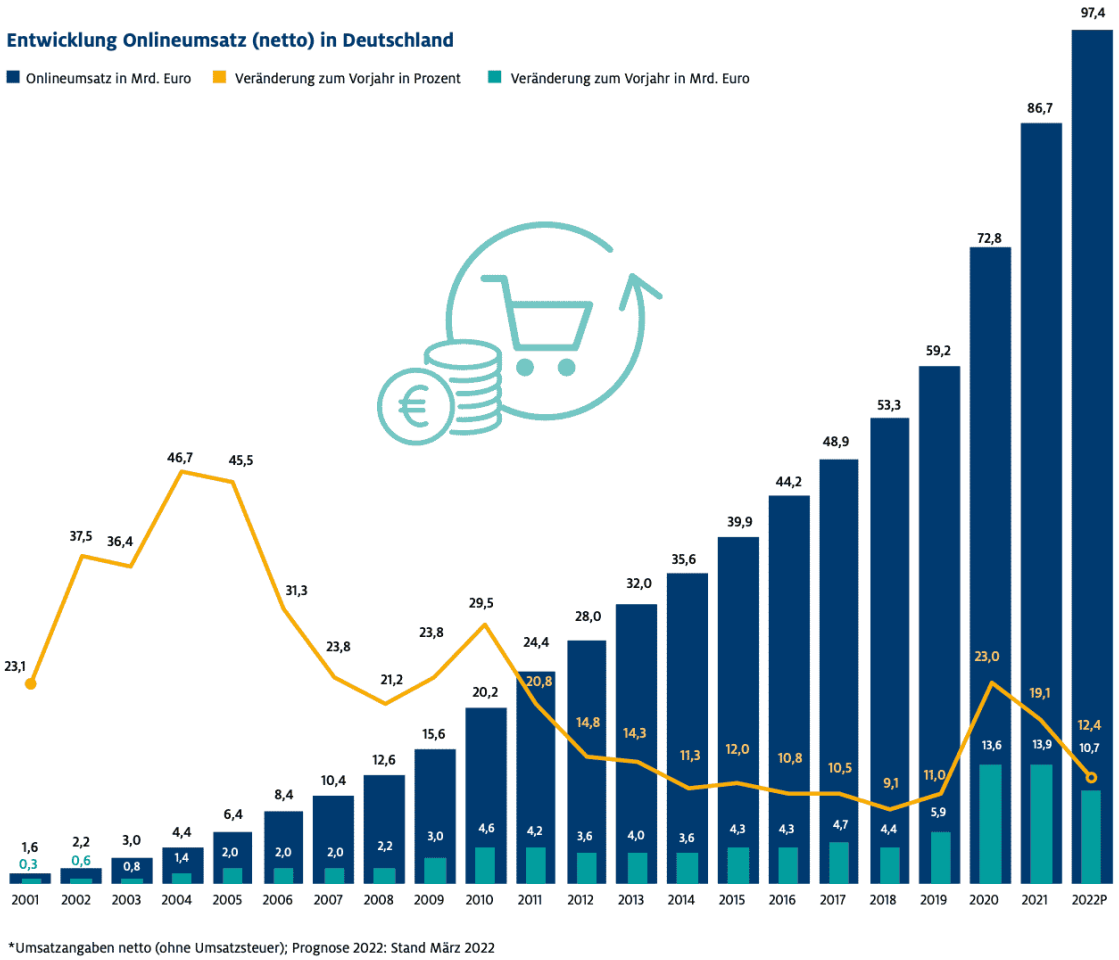


Abbildung 1.1: Umsatzentwicklung E-Commerce von 2001-2022

Mithilfe von Machine Learning können aus vorhandenen Daten, Nutzergruppe bestimmt werden, mit deren Hilfe besser Angebote für die Nutzergruppen erstellt und ausgespielt werden können.

Allgemein benennt [talend.com] folgende drei Vorteile für Unternehmen der durch den Einsatz von maschinellen Lernen entstehen.

1. Future-Proof: es ist eine Technologie der Zukunft, die in den kommenden Jahren weiter entwickelt wird.

2. Effizienz: durch die Automatisierung und Klassifizierung lassen sich große Datenmengen effizienter analysieren. Dadurch verbessern sich interne und externe Unternehmensprozesse.

3. Fördert Konzentration: ML übernimmt repetitive Aufgaben, so bleibt Mitarbeitern für anspruchsvollere Arbeiten.

## 1.1 Motivation

Im E-Commerce nutzen Onlineshop Betreiber Kundenklassifizierungen, um Kunden einzuteilen und ihnen besser Angebote unterbreiten zu können. Oft sind diese Klassifizierungen in den ersten Schritten bei der Shop-Planung oder bei einem Relaunch erstellt wurden, basierend auf einer erwarteten und anvisierten Kundengruppe. In vielen Fällen werden keine Kundenklassifizierungen erstellt. Es wird unterstellt das alle Menschen zur Zielgruppe gehören.

### 1.1.1 Ausgangslage

Im Rahmen dieser Arbeit soll ein kleiner Onlineshop eines Verlages aus dem Nordosten Deutschlands untersucht werden. Über den Shop werden hauptsächlich Print-Produkte wie Bücher und Zeitschriften. Eine vorherige Planung und Analyse der Nutzergruppen und ihrer Bedürfnisse sind einigen Mitarbeitern, die den Shop betreuen, nicht wichtig.

Folgende Aussagen trafen Mitarbeitern, die unmittelbar mit den Onlineshop Arbeiten und diesen betreuen.„Für den Onlineshop benötigen wir keine Nutzergruppen, dafür haben wir die verschiedene Menüpunkte. Nutzergruppen werden erst interessant, wenn wir auf verschiedenen Kanäle Werbung schalten.“ Aussage einer Mitarbeiterin für Kundenmanagement. „Zielgruppendefinition ist wichtig, aber für den Shop sollte keine angefertigt werden.“ Aussage einer Mitarbeiterin des Webdesigner Teams.

### 1.1.2 Vorteile durch die Vorhersage

Einer der wichtigsten Vorteile einer durch künstlicher Intelligenz vorhergesagten Nutzergruppenklassifizierung wäre das Ausschließen des menschlichen Einwirkens. Das Erhöht die Genauigkeit der Entscheidungen und es passieren weniger Fehler. Des Weiteren erkennt künstliche Intelligenz versteckte Muster aus tausenden und mehr Datensätzen. Ein weiterer Vorteil ist die Möglichkeit einer zeitnahen Anpassung der Verkaufsprozesse auf der Webseite und eine kontinuierliche Optimierung des Onlineauftritts.

## 1.2 Ziel der Arbeit

In dieser Arbeit ist die Prüfung, ob eine mögliche Vorhersage der Nutzergruppenklassifizierung durch künstliche Intelligenz möglich ist und ob diese Vorhersage Vorteile

gegenüber einer klassischen Klassifizierung, beispielsweise durch Personas ist. Für die Vorhersage werden vorhandene Daten aus dem Shop und verschiedene externe Quellen verwendet.

Durch die Zielsetzung lässt sich folgende Forschungsfrage und dazugehörige Unterfragen formulieren:

„Ist die Erstellung eine Nutzergruppenklassifizierung mit Machine-Learning besser als eine klassische Analyse der Nutzergruppen?“

- Welche Systeme und Algorithmen eignen sich am besten für die Klassifizierung?

## 1.3 Inhaltlicher Aufbau der Arbeit

Diesen Teil schreibe ich während der Fertigstellung der einzelnen Kapitel.

# Kapitel 2

# Grundlagen

## 2.1 Grundbegriffe

### Big Data

Eine Abgrenzung des Begriff Big Data ist nicht eindeutig, da der Begriff sehr heterogen verwendet wird. Somit gibt es viele Definitionen vom Begriff Big Data.

Als Big Data werden Daten bezeichnet, die entweder zu groß, zu komplex, zu schnelllebig oder zu schwach strukturiert sind, um diese mit herkömmlichen Methoden auszuwerten. Big bezieht sich in der Definition auf die vier Dimensionen. In Big Data sind Technologien die richtigen Informationen dem richtigen Adressaten zur richtigen Zeit in der richtigen Menge am richtigen Ort und in der erforderlichen Qualität bereitstellen.

Auch strukturierte Daten. Verschiedene (autonome) Datenquellen (Datenbanken oder Anwendungen).

Auf volume (Umfang, Datenvolumen), velocity (Geschwindigkeit, mit der die Datenmengen generiert und transferiert werden), variety (Bandbreite der Datentypen und -quellen) und veracity (Echtheit von Daten).

### Knowledge Discovery in Data (-base)

Dies hat das Ziel aus vorhandenen meist großen Datenbeständen, fachliche Zusammenhänge zu erkennen. Zu den Teilschritten des KDD Prozesses gehören 1. Bereitstellung von Hintergrundwissen, 2. Definition der Ziele, 3. Datenauswahl, 4. Datenbereinigung, 5. Datenreduktion, 6 Auswahl eines Modells, 7. Data-Mining, die eigentliche Datenanalyse, 8. Interpretation der gewonnenen Erkenntnisse.

### Data-Mining

Data-Mining die systematische Anwendung von statischen Methoden auf große Datenbestände, um neue Querverbindungen zu erkennen. Data-Mining ist ein Teilprozess KDD Prozesses. Mit Data-Mining findet ein Informationsgewinn oder -erweiterung, aus den Big Data statt. Hier werden Algorithmen aus dem Bereich ML angewandt.
k-Means Algorithmus

### Machine Learning

Maschinelles Lernen (eng. Machine Learning) ist nach [https://datasolut.com/was-ist-

machine-learning] ein Teilbereich der künstlichen Intelligenz, der System in die Lage versetzt, automatisch aus Erfahrungen (Daten) zu lernen und sich zu verbessern. Aufgaben die Machine Learning erledigen kann, ist Berechnung von Wahrscheinlichkeiten für bestimmte Ereignisse, Erkennen von Gruppen und Clustern in Datensätzen, Erkennen von Zusammenhängen in Sequenzen, Reduktion von Dimensionen ohne großen Informationsverlust und Optimierung von Geschäftsprozessen.

**Künstliche Intelligenz**

Eine einheitliche gemeingültige Definition von künstlicher Intelligenz zu geben ist nicht einfach. Zuvor muss Intelligenz definiert werden. Aber was ist Intelligenz? In der Literatur werden kognitive Fähigkeiten oft mit Intelligenz in Verbindung gebracht. In der Definition von künstlicher Intelligenz gibt es schwache oder enge künstliche Intelligenz, die auf die Lösung bestimmter Aufgaben beschränkt ist und menschliche Intelligenz nicht imitieren kann. Starke oder allgemeine künstliche Intelligenz hingegen ist in der Lage die kognitiven Fähigkeiten des Menschen zu erzielen.

Weitere Definitionen werden im Verlauf der Arbeit ergänzt.

## 2.2 Verwandte Arbeiten

Mit weiteren Anwendungsfällen beschäftigt sich das ,,Fraunhofer Institut BIG DATA" in ihrem Paper [Big Data Frauenhofer].

Um den Einsatz künstlicher Intelligenz im werte-orientierten Marketing zu bewerten, befassten sich in [EConster] einige Professoren des ,,Leibniz-Informationszentrum Wirtschaft" mit diesem Thema.

Die Masterarbeit [Bitstream] von Eduard Weigandt befasst sich mit der Personalisierung im E-Commerce basierend auf Data-Mining. Interessante Grundlagen zum Einstieg in E-Commerce und künstlicher Intelligenz sind auf [Eqop] zu finden. Die Firma Kobold AI befasst sich in ihrem Artikel [Kobold.ai] ,,Optimale Segmentierung von Bestandskunden durch KI" und erläutert Methoden zur Clustering von Bestandskunden. ,,Datasolut" ist ein weiteres Unternehmen das sich in [Datasolut1] mit Kundenklassifizierung, Clusteranalyse und maschinellem Lernen befasst. Ebenfalls von ,,Datasolut" ist der Artikel [Datasolut2] in dem erfolgreiche Anwendungen und Beispiele zum Thema künstlicher Intelligenz im E-Commerce aufgezeigt werden.

## 2.3 Big Data

Big Data beschäftigt sich nach [DataSolut3] und [Oracle] mit dem Sammeln, Verarbeiten und Zusammenführen von großen Datenmengen. Um diese Daten für die Entscheidungsfindung und Prozessautomatisierung zu verwenden. Dabei stammen die Daten aus den unterschiedlichsten Quellen, aus verschiedenen Datenbanken oder auch direkt aus Programmen. Als Datenquellen können folgende infrage kommen:

- Internetnutzung
- Social Media

- Geo-Tracking
- Cloud Computing
- Vitaldaten-Messung
- Media-Streaming

Diese Daten können strukturiert, aber auch unstrukturiert vorliegen. [Gratner] beschrieb Big Data anhand von den ,,4 V's". Mit der Zeit wurde es um ein ,,V" erweitert. Diese Beschreibung wird in unterschiedlichen Publikationen aufgegriffen, unter anderen auch in [Oracle].

### Volume (Volumen)

Immer größere Datenmengen müssen Verarbeitet werden. Durch die stetig zunehmende Digitalisierung in immer mehr Lebensbereichen wächst die erzeugte Datenmenge pro Zeiteinheit immer mehr an. So werden großen Datenmengen nicht nur durch die oben genannten Quellen erzeugt, sondern auch z. B. durch Gerätesensoren. Hierbei können etliche Terabytes oder hunderte Petabytes an Daten anfallen. Wie die Abbildung 2.1 [Statistika] zeigt, wird das Datenvolumen im Jahr 2025 auf 181 Zettabyte vorhergesagt.

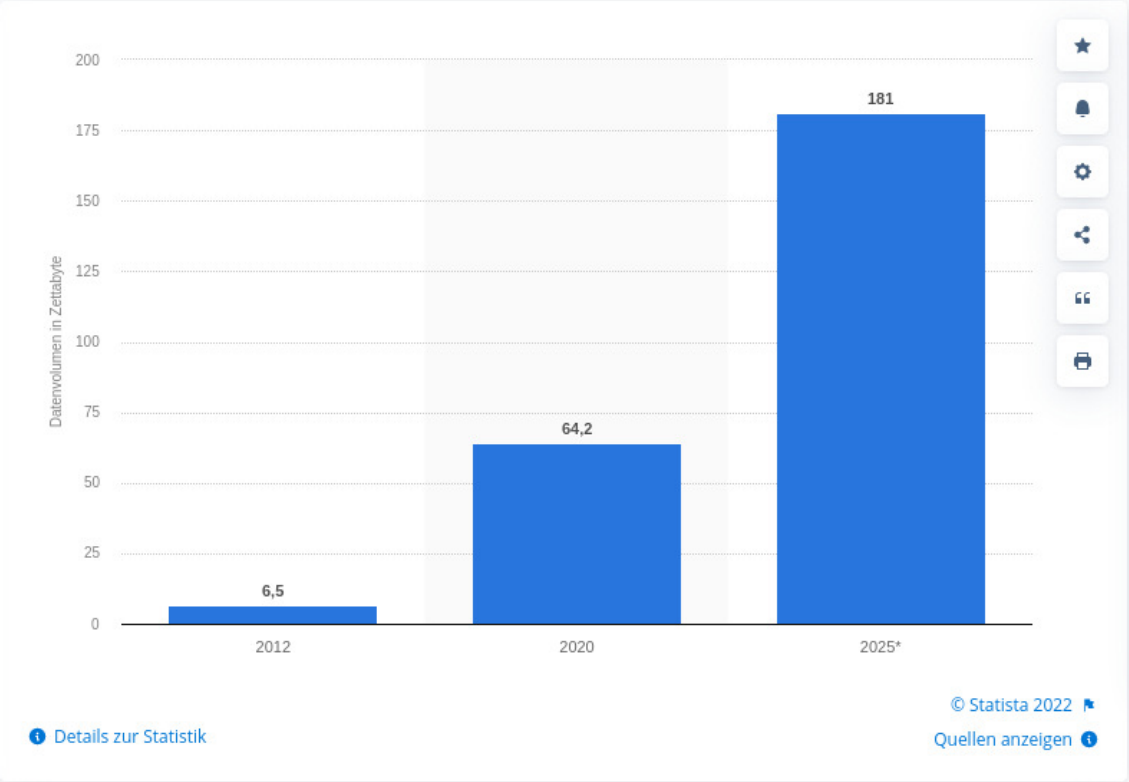


Abbildung 2.1: Umsatzprognose E-Commerce bis 2025

### Variety (Vielfalt)

Durch die unterschiedlichen Bereiche, in denen die Datenmengen entstehen sind, diese sehr unterschiedlich und zu meist unstrukturiert. Oft liegen diese in relationalen Datenbanken und können dort nicht ausgewertet werden. Neben Texten liegen die Daten in Bildern und Videos vor die Analyse erfolgt durch Machine Learning Algorithmen .

#### Velocity (Geschwindigkeit)

Mit der Entwicklung der Technik produzieren Softwaresysteme mit einer höheren Geschwindigkeit mehr Daten. Bei vielen Produkten fließen die Daten nicht auf eine Festplatte, sondern werden direkt im Speicher verarbeitet. Solche Produkte arbeiten in Echzeit oder beinahe in Echtzeit. Deren Verarbeitung in immer kürzerer Zeit erfolgt. Für Unternehmen und verschiedenen Use Cases kann die Verarbeitung in Echtzeit einen erheblichen Wettbewerbsvorteil bedeuten.

#### Veracity (Wahrhaftigkeit)

Da die Daten oft aus Quellen kommen, deren Wahrheitsgehalt nicht sicher ist und die Daten in nicht geeigneter Qualität vorliegen, können diese nicht ohne eine aufwendige Nachbearbeitung eingesetzt werden.

#### Value (Mehrwert)

Durch die Verknüpfung der Daten, die beim Einsatz der Techniken des Machine Learning entstehen, ist dieser Mehrwert eines der wichtigsten ,,V" bei Big Data. Ohne diesen Mehrwert würde Big Data keinen Sinn ergeben.

### 2.3.1 Technologien

Rund um das Thema Big Data, haben sich verschiedene Technologien entwickelt, die Ansätze für die Verarbeitung von großen Datenmengen liefern. Nachfolgende sind Open Source Produkte stellvertretend einige genannt.

#### Apache Spark

Es ist nach eigenen Angaben [Spark] eine mehrsprachige Engine zur Ausführung von Data Engineering, Data Science und maschinelles Lernen auf Single-Node-Maschinen oder Clustern.

#### Apache Hadoop

Hadoop ist nach [Hadoop] ein Framework das mit einfachen Programmiermodellen, welches eine verteilte Verarbeitung von großen Datenmengen anbietet, das von einzelnen Server auf mehrere tausend skaliert werden kann.

#### Apache Cassandra

Nach eigener Beschreibung [Cassandra], ist Cassandra ist ein skalierbares und hochverfügbares verteiltes Datenbanksystem. Es basiert auf NoSQL, ist Open-Source und kann ebenfalls auf einzelnen Servern oder in der Cloud eingesetzt werden.

## 2.4 Data-Mining

## 2.5 Clustering

Abkürzungen
ML Mashine Learning
KI Künstliche Intelligenz

# Kapitel 3

# Explorative Datenanalyse

Das Wichtigste beim maschinellen Lernen sind die Daten. Beim Erheben der Daten für maschinelles Lernen sollten die Datensätze nicht aus kontrollierten Umgebungen stammen die mit festgelegten Rahmenbedingungen arbeiten. Sonst können die Algorithmen unter realen Bedingungen schlecht abschneiden, da hier andere Daten von den Trainingsdaten zu stark abweichen. Daher sollte man die Daten kennen die verwendet werden sollen.

Die explorative Datenanalyse (EDA) wird verwendet, um Daten zu analysieren und ihre Merkmale zusammenzufassen und um besseres Verständnis für die Datensätze zu bekommen. Dies hilft dabei, um herauszufinden wie die Daten am besten verarbeitet werden können. Dabei hilft die EDA, Muster oder Fehler und Anomalien zu finden, sowie Hypothesen testen und Annahmen zu überprüfen.

Sind die analysiert können sie mithilfe, beispielsweise von maschinelles Lernen verarbeitet werden.

## 3.1 Arten der EDA

Primär unterschiedet [ibm.com] vier Arten der EDA.

**Univariat, nicht-grafisch**
Untersucht nur eine Variable und ist somit die einfachste Form der Datenanalyse. Hierbei geht es nicht um Ursachen Forschung oder Finden von Beziehungen, sondern das Beschreiben der Daten und Muster zu finden.

**Univariat, grafisch**
In dieser Art gibt es beispielsweise die Möglichkeit eine Variable beispielsweise in einem

- Stamm-Blatt-Kurvendiagramm[1],
- Histogramm[2] oder
- Box-Diagramm[3] dargestellt.

[1]Stamm-Blatt-Kurvendiagramm zeigt die alle Datenwerte und die Form der Verteilung.
[2]Histogramm mit Balken wird die Häufigkeit oder Anteil der Fälle für einen Wertebereich angezeigt.
[3]Box-Diagramm fünfstellige Zusammenfassung von 1. Minimum, 2. erstes Quartils, 3. Median, 4. dritten Quartils, 5. Maximum.

**Multivariat, nicht-grafisch**
Diese Daten bestehen aus mehreren Variablen. Sie zeigen allgemeine Beziehungen zwischen zwei oder mehreren Variablen durch Kreuztabellen oder Statistiken.

**Multivariat, grafisch**
Zeigen grafisch die Beziehungen zwischen ein oder mehreren Variablen. Häufig werden zur Darstellung Streu-[4], Lauf-[5], Blasendiagramme[6], Multivariate Diagramme[7] und Heat-Maps[8] verwendet.

## 3.2 Methoden der Datenanalyse

## 3.3 Kundengruppen

### 3.3.1 DiSG Modell

Wichtig das DISG-Modell (dominant, initiativ, stetig, gewissenhaft)

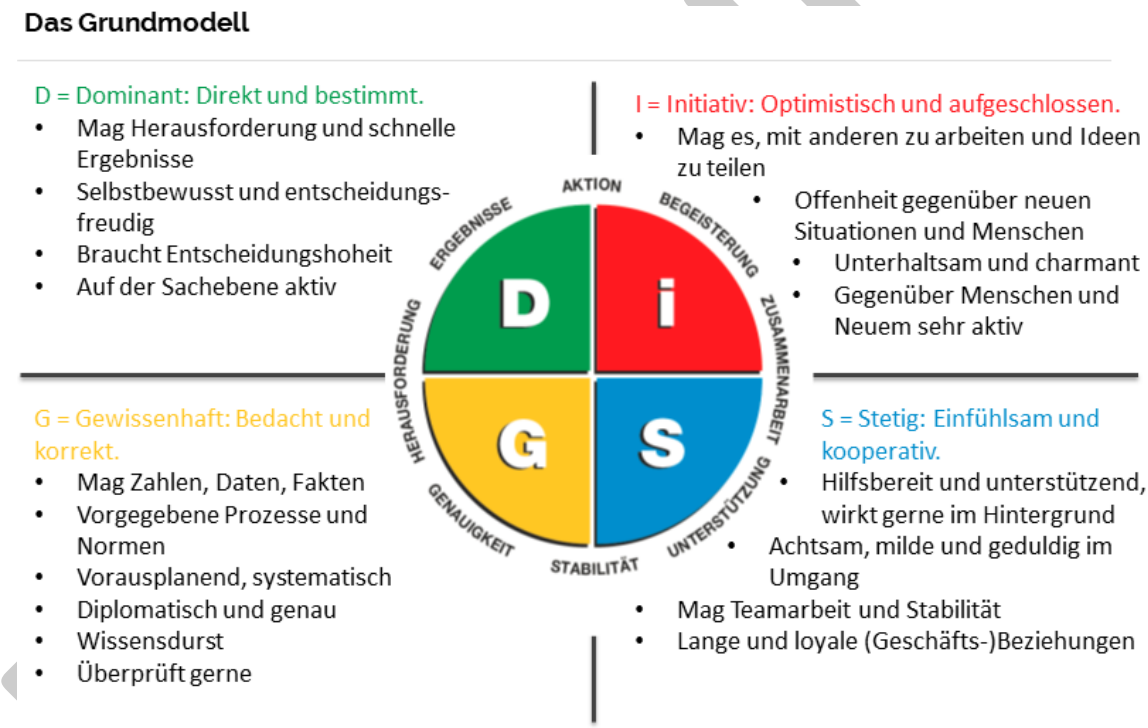


Abbildung 3.1: DiSG Übersicht https://www.disg-modell.de

Der *Dominante*, hat konkrete Erwartungen an ein Produkt. Daher sollte eine Produktbeschreibung nicht nur klar und deutlich sein, sonder ebenso dessen Nutzen, welche Funktionen es hat, warum es so gut funktioniert und die wie es die Probleme des Kunden löst.

[4]Streudiagramm, stellen Datenpunkte auf einer horizontalen und einer vertikalen Achse, die Abhängigkeit einer Variablen zu einer anderen zu zeigen.
[5]Laufdiagramm ist ein Liniendiagramm von Daten die über die Zeit aufgetragen werden.
[6]Blasendiagramm ist eine Datenvisualisierung mittels Kreisen.
[7]Multivariate Diagramm ist eine grafische Darstellung zwischen Faktoren und einer Antwort.
[8]Heat-Map hierbei werden die Daten durch Farben dargestellt.

Der *Initiative* Kundentyp ist begeisterungsfähig, extrovertiert und optimistisch. Da dieser Typ eher die positiven Eigenschaften eines Produktes sieht, lohnt es sich diesen Kundentyp Rezensionen schreiben zu lassen. Er legt Wert auf eine Wertbeschreibung des Produktes. Dies kann mit Storytelling erfolgen.

Der *Stetige* kann eine harte Nuss für Onlinehändler sein, da dieser Kundentyp sehr viel Wert auf Produktbeschreibung legt, die dessen Bedürfnisse und Ziele erfasst. Andererseits ändern sich die Bedürfnisse dieses Typs nicht über Nacht. Somit kann er zu einem regelmäßigen kaufenden Stammkunden werden. Die Produktbeschreibung sollte die Unique Selling Point enthalten und herausstellen welche Ziele das Produkt, wie unterstützt.

Der *Gewissenhafte* geht systematisch vor und analysiert seine Erkenntnisse zum Produkt. Ist eher reserviert und zurückhaltend und schreibt somit weniger Rezensionen. Dieser Typ kann mit Daten und Fakten überzeugt werden. Dies sollte mit Anwendungsbeispielen und Studienergebnissen untermauert werden.

### 3.3.2 Mögliche Kundengruppen

Nach KI im Onlineshop: So steigern Onlinehändler ihren Umsatz

Der *Performer* (hat ernste Kaufabsicht und weiß genau was er will) nutzt Suchfunktion und geht gleich in die Kategorie. ≪Diesem Kunden kann die Suchfunktion besonders angezeigt werden, da er zielstrebig etwas sucht. Findet er es nicht, sucht er auf anderen Seiten weiter.≫

Der *Stöberer* (will sich beim stöbern inspirieren lassen) klick auf der Startseite viel herum. ≪Diesem Kunden könnten Angebote angezeigt werden, die diesem Kunden zum Kauf anregen.≫

Nach Die 6 häufigsten Kundentypen und wie man sie am besten anspricht

Der *Schnäppchenjäger* (Hautsache Preiswert) dieser Kunde baut keine richtige Beziehung zu einem Onlineshop auf. Diese Kunden besuchen meist mehrere Webshops und/oder kommen von Preissuchmaschinen. Diesen Kunden zu binden ist eine Herausforderung. ≪Dieser Kunde kann effektiv mit Rabatte, Gutscheincodes oder Ausverkaufsaktionen ansprechen. Auch E-Mail-Aktionen lassen sich bei diesem Kundentyp leicht anbringen.≫

Der *Eilige* (stehen unter Zeitdruck oder wollen sie sich nicht nehmen) bringt keine Zeit für ein ausgiebiges Shoppingerlebnis mit. Sie verfügen oft über hohe monetäre Mittel, um sich beispielsweise Expresslieferung zu leisten.≪Er braucht eine strukturierte Webseite, auf die er Informationen schnell findet. Dies kann durch Suchfunktion für Preis, Größe, usw., eine Schnellansicht für Artikel auf der Übersichtseite mit deutlichem Produktabbildungen.≫

Der *Stöberer* (sind Gelegenheitskäufer) schaut stundenlang im Internet ohne bestimmte Absicht. Hierbei ist es schwer vorhersagbar, was dieser Kunde kauft. Oft legt dieser Kunde ein oder zwei Artikel in den Warenkorb, schließt die Bestellung nie ab.≪Diese Kunden kennen die Preise im Internet und lassen sich evtl. durch ein Rabattangebot oder Preisnachlass umstimmen.≫

Der *Sammler* (auf der Suche nach Besonderheiten) sucht Unikaten oder ganz bestimmte Artikel wie Sondereditionen oder limitierte Ausgaben. Lieferzeiten und Preise spielen keine Rolle. ≪Bei diesem Kunden sollten Lagerinformationen stets aktuell sein. Ist nach

dem Bestellvorgang ersichtlich, dass der Artikel nicht mehr vorhanden ist, wird meist storniert und der Kunde kommt nicht mehr wieder.≫

Der *Misstrauische* (Angst Oper zu werden) ziehen große bekannte Onlinehändler vor, da sie befürchten das bei kleineren Probleme auftreten können. Das Misstrauen geht so weit, das dieser Kundentyp nur per Nachname bestellt oder die Ware Selbst abholt. ≪Dieser Zielgruppe sind Rückgaberecht und Garantieleistungen besonders wichtig. Ebenfalls können diese Kunden durch zusätzliche Versicherungen gewonnen werden.≫

Der *Schüchterne* (anonym ist wichtig) kaufen online, da es ihnen peinlich ist, den gesuchten Artikel im stationären Handel zu kaufen. ≪Diese Kunden nutzen häufig Suchmaschinen und gelangen so in den Shop (SEO ist hier wichtig). Ein weiteres Kriterium für den Schüchternen sind seriöse Bezahl- und Versandarten.≫

## 3.4 Daten für die Onlineshop-Nutzeranalyse

Hier auch noch etwas über Datenintegration.

# Kapitel 4

# Maschinelles Lernen

Das maschinelle Lernen ist ein Teilgebiet der künstlichen Intelligenz und übernimmt Aufgaben die typischerweise menschliche Intelligenz erfordern. Sie soll dabei helfen Muster und Gesetzmäßigkeiten in Datensätzen zu erkennen. Aus vorhandenen Daten wird durch Algorithmen künstliches Wissen generiert.

## 4.1 Methoden maschinellen Lernens

Maschinelles Lernen lässt sich nach [talend.com] in vier Methoden unterteilen. Die in den folgenden Kapiteln näher betrachtet werden.

### 4.1.1 Überwachtes Lernen

Bei überwachtem Lernen erhält ein Computer strukturierte Inputs und gewünschte Ergebnisse. Nun muss der Computer Wege finden mit Inputs, um diese Ergebnisse zu erreichen, d.h der Algorithmus versucht eine vorhersage Funktion zu entwickeln. Die Vorhersagen über die unbekannten oder künftigen Daten wird als prädiktive Modellierung bezeichnet.

Das überwachte Lernen lässt sich in zwei Arten unterteilen.

- Klassifizierung: das Ergebnis ist eine Kategorie, z. B. Gruppenzugehörigkeit
- Regression: hier ist das Ergebnis ein realer Wert, z. B. Produktpreis

Mittels verschiedener Methoden lassen sich die Ergebnisse vorhersagen. Diese können Entscheidungsbäume, Random-Forest-Algorithmus, lineare Regression, Naive-Bayes-Verfahren, usw. Eignet sich für Probleme der Klassifizierung und Regression.

### 4.1.2 Unüberwachtes Lernen

Bei diesem Lernen sind keine strukturieren Daten vorhanden, eher legen sie unstrukturiert und unbeschriftet vor. Der Algorithmus muss die Strukturen selbst erkennen. Aus erkannten Mustern und Merkmalen, lassen sich weitere Muster und Korrelationen vorhersagen. Ist, gibt zwei Arten von unüberwachten Lernen.

- Clustering: Gruppierung von Daten, weitere lassen sich in bestehende Cluster zuordnen
- Assoziation: Regeln in Daten finden, so werden Daten durch Erfahrung definiert

Zu diesem Lernen gehören u. a. K-Means, hierarchische Clusteranalyse und Dimensionsreduktion. Mit dieser Methode lassen sich Probleme des Clustering, Dimensionsreduktion und Lernen von Assoziationsregeln.

### 4.1.3 Teilweise überwachtes Lernen

Es ist ein Hybridverfahren zwischen unüberwachten und überwachten Lernen. Die Rohdaten sind nur teilweise strukturiert und beschriftet. Durch die strukturierten Daten werden die unstrukturierten aufgewertet.

Die strukturierten Daten finden anfangs Verwendung, um diese auf Muster und Korrelationen zu untersuchen. Im Anschluss können diese auf die unstrukturierten angewandt werden. Mit dem teilweise überwachten Lernen können Probleme der Klassifizierung und Regression gelöst werden.

### 4.1.4 Bestärktes Lernen

Ein Computerprogramm interagiert mit einer dynamischen Umgebung. Beim Ausführen bestimmter Aufgaben erhält das Programm gutes oder schlechtes Feedback für die Aktion. Durch die Belohnung und Bestrafung lernt das Programm die richtigen Verhaltensweisen. Belohnungen werden auf zwei verschiedene Arten vergeben.

- Monte Carlo: Vergabe erfolgt am Ende
- Temporal-Difference-Learning (TD-Learning): Vergabe der Belohnung erfolgt nach jedem Schritt

Als Algorithmen sind hier beispielsweise Q-Learning, Deep Q Network (DQN) und State-Action-Reward-State-Action (SARSA) zu nennen.

## 4.2 Algorithmen zum Analysieren von Onlineshop Daten

Die verwendeten Algorithmen des maschinellen Lernens können durch ihre Ähnlichkeiten gruppiert werden. In den folgenden Kapiteln werden diese wie in [machinelearningmastery.com] diskutiert.

### 4.2.1 Regressionsalgorithmen

(eng. Regression Algorithms)

Durch die interaktive Verwendung eines Fehlmaßes erfolgt eine Verfeinerung des Algorithmus, der Regressionsalgorithmen (eng. Regression Algorithms) befasst sich mit den Beziehungen zwischen Variablen.

Wichtige Algorithmen

- Ordinary Least Squares Regression (OLSR)
- Linear Regression

- Logistic Regression
- Stepwise Regression
- Multivariate Adaptive Regression Splines (MARS)
- Locally Estimated Scatterplot Smoothing (LOESS)

### 4.2.2 Instanz-basierende Algorithmen

(Instance-based Algorithms)
Algorithmen

- k-Nearest Neighbor (kNN)
- Learning Vector Quantization (LVQ)
- Self-Organizing Map (SOM)
- Locally Weighted Learning (LWL)
- Support Vector Machines (SVM)

### 4.2.3 Regularisierungsalgorithmen

(eng. Regularization Algorithms)
Algorithmen

- Ridge Regression
- Least Absolute Shrinkage and Selection Operator (LASSO)
- Elastic Net
- Least-Angle Regression (LARS)

### 4.2.4 Entscheidungsbaumalgorithmen

(eng. Decision Tree Algorithms)
Algorithmen

- Classification and Regression Tree (CART)
- Iterative Dichotomiser 3 (ID3)
- C4.5 and C5.0 (different versions of a powerful approach)
- Chi-squared Automatic Interaction Detection (CHAID)
- Decision Stump
- M5
- Conditional Decision Trees

### 4.2.5 Bayessche Algorithmen

(eng. Bayesian Algorithms)

Algorithmen

- Gaussian Naive Bayes
- Multinomial Naive Bayes
- Averaged One-Dependence Estimators (AODE)
- Bayesian Belief Network (BBN)
- Bayesian Network (BN)

### 4.2.6 Clustering-Algorithmen

(eng. Clustering Algorithms)

Algorithmen

- k-Means
- k-Medians
- Expectation Maximisation (EM)
- Hierarchical Clustering

### 4.2.7 Lernalgorithmen für Assoziationsregeln

(eng. Association Rule Learning Algorithms)

Algorithmen

- Apriori algorithm
- Eclat algorithm

### 4.2.8 Neuronale Netze

(eng. Artificial Neural Network Algorithms)

Algorithmen

- Perceptron
- Multilayer Perceptrons (MLP)
- Back-Propagation
- Stochastic Gradient Descent
- Hopfield Network
- Radial Basis Function Network (RBFN)

### 4.2.9 Depp-Learning-Algorithmen

(eng Deep Learning Algorithms)

Algorithmen

- Convolutional Neural Network (CNN)
- Recurrent Neural Networks (RNNs)
- Long Short-Term Memory Networks (LSTMs)
- Stacked Auto-Encoders
- Deep Boltzmann Machine (DBM)
- Deep Belief Networks (DBN)

### 4.2.10 Dimensionsreduktionsalgorithmen

(eng. Dimensionality Reduction Algorithms)

Algorithmen

- Principal Component Analysis (PCA)
- Principal Component Regression (PCR)
- Partial Least Squares Regression (PLSR)
- Sammon Mapping
- Multidimensional Scaling (MDS)
- Projection Pursuit
- Linear Discriminant Analysis (LDA)
- Mixture Discriminant Analysis (MDA)
- Quadratic Discriminant Analysis (QDA)
- Flexible Discriminant Analysis (FDA)

### 4.2.11 Ensemble-Algorithmen

(eng. Ensemble Algorithms)

Algorithmen

- Boosting
- Bootstrapped Aggregation (Bagging)
- AdaBoost
- Weighted Average (Blending)

- Stacked Generalization (Stacking)
- Gradient Boosting Machines (GBM)
- Gradient Boosted Regression Trees (GBRT)
- Random Forest

## 4.3 Umsetzung der Datenverarbeitung

Erste Tests in Sachen Python und Datenanalyse.

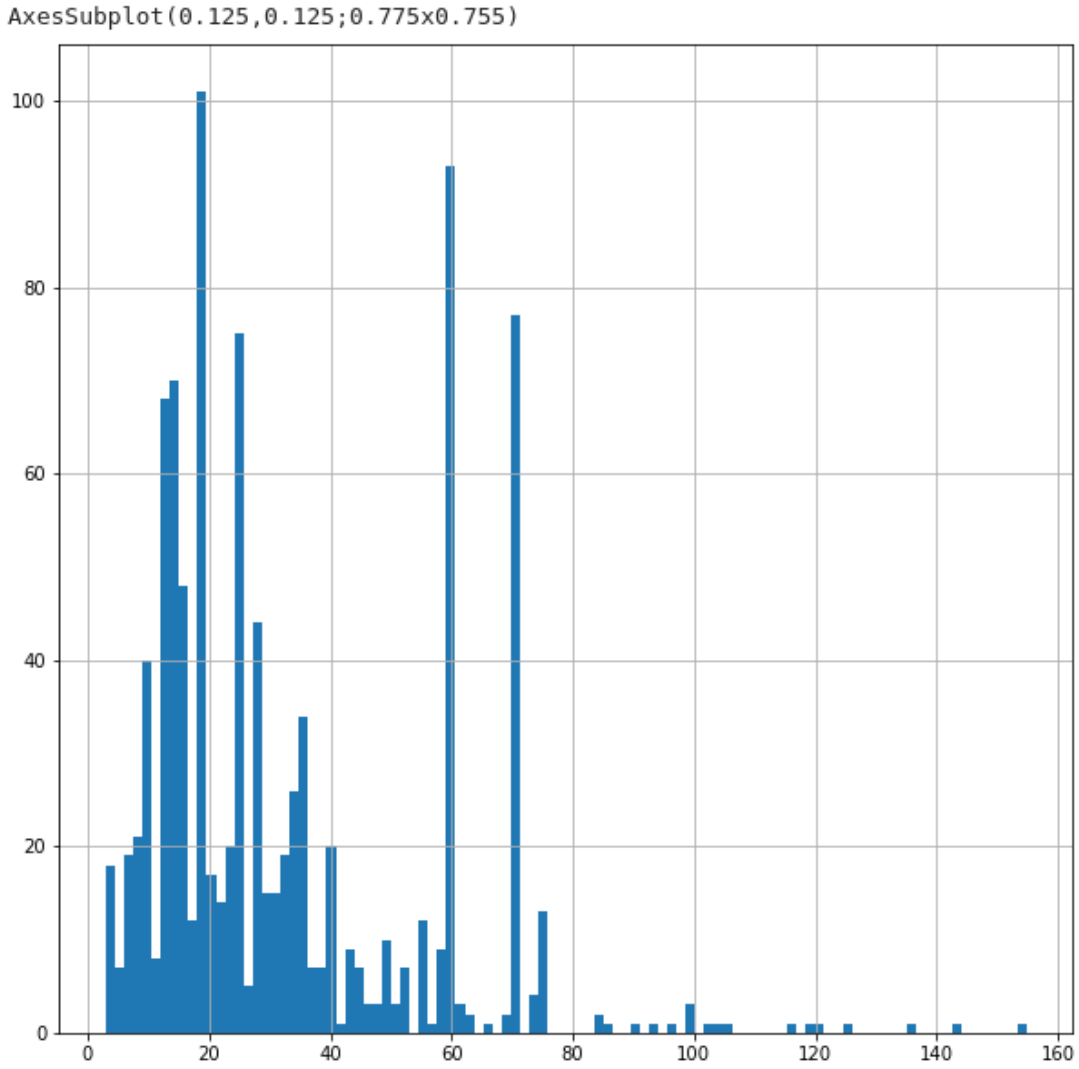


Abbildung 4.1: Anzahl der Bestellung in Abhängigkeit der Warenkorbhöhe

## 4.4 Umsetzung der Clusterung

# Kapitel 5

# Evaluation

## 5.1 Aufbau der Umgebung

Hier könnte das CMS erwähnt werden.

## 5.2 Ergebnisse

## 5.3 Bewertung und Diskussion

# Kapitel 6

# Future Work

# Anhang

# Literaturverzeichnis

[1] P. Versteegen, "Anlagetrend E-Commerce Aktien: Die besten E-Commerce-Wertpapiere." Finanzwissen, Aug. 2022. [Online]. https://finanzwissen.de/aktien/e-commerce/. [abgerufen am 21.11.2022].

# Glossar

**CMS** Nutzerfreudliche Bedienungsoberfläche einer Software. 9