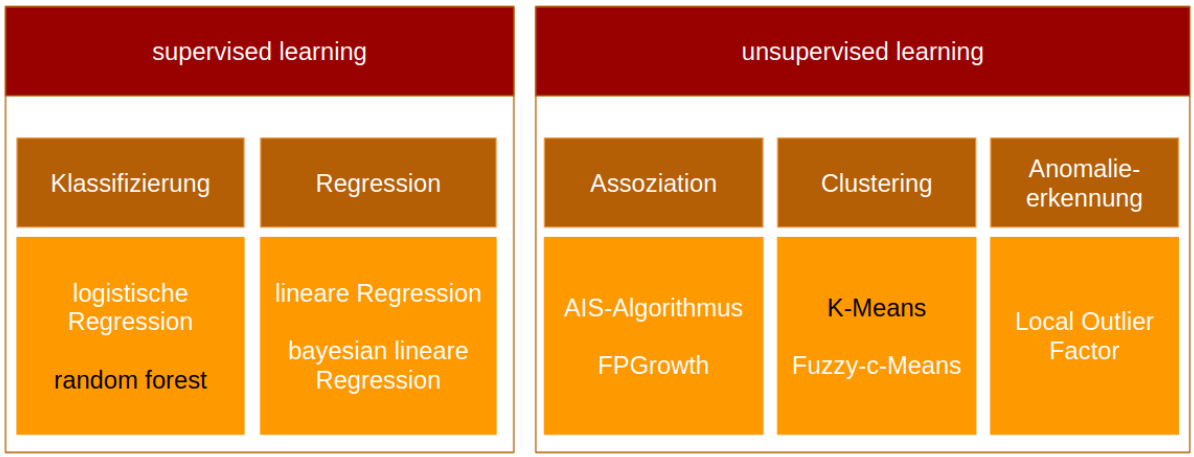
IT-Seminar

im Studiengang

BSc Computer Science

Data Mining

Einführung und Beispiele



Verfasser : Adrian Berger
adrian.berger.2@students.bfh.ch

Dozent : Jürgen Eckerle

Modul : BTI7311

Datum : 25.05.2021

Abstract

Diese Arbeit bietet einen Überblick über die Thematiken im Bereich Data Mining. Dazu werden einleitend Beispiele vorgestellt, bei welchen Data Mining verwendet wurde. Hierbei handelt es sich einerseits um Beispiele aus dem Bereich Gesundheit, Sicherheit, Landwirtschaft sowie ein kommerzielles Beispiel. Zudem wird in der Arbeit eine kurze Übersicht zu den ethischen Überlegungen und Bedenken gegeben. Der Hauptteil der Arbeit beschäftigt sich mit den verschiedenen Verfahren im Bereich Data Mining und Machine Learning.

Dabei soll eine Übersicht über die Verfahren, sowie die wichtigsten Stichwörter dazu gegeben werden. Diese Übersicht soll es der Leserschaft ermöglichen einen Überblick zu den Verfahren zu erhalten, aber auch gezielt weiterführende Themen ansprechen. Somit soll die Leserschaft gezielt motiviert werden, die weiterführenden Themen nach eigenen Interessen durch die referenzierten Dokumente zu vertiefen.

Gegen Ende des Dokumentes werden konkret die Verfahren Random Forest und K-means vorgestellt. Diese dienen beispielhaft dazu, die Komplexität des Data Mining aufzuzeigen. Auch hier werden gezielt Stichwörter eingesetzt, welche die Leserschaft für eigene, weitere Recherchen verwenden kann.

Inhaltsverzeichnis

1 Einleitung

1.1 Definition

Data Mining ist ein Sammelbegriff für die Anwendung statistischer Verfahren auf grosse Datenmengen. Das Ziel dieser Verfahren ist es, aus den Daten neue Erkenntnisse zu gewinnen, wie etwa Korrelationen oder Trends zu identifizieren. Durch die Anwendung von Data Mining können in grossen Datenmengen Muster und Regeln erkannt werden, die durch manuelle Auswertungen nur schwer zu erkennen sind. Entgegen der möglichen Erwartung aus dem Begriff Data Mining, geht es bei diesen Verfahren nicht darum, Daten zu generieren oder zu sammeln. Dieser Prozessschritt geschieht jeweils vor dem Data Mining und ist somit nicht Teil vom Data Mining Prozess. [1]

1.2 Erläuterungen

Grundsätzlich beruht Data Mining meist auf Verfahren der Statistik und somit aufgrund mathematischer Grundlagen. In vielen Fällen beruhen diese statistischen Verfahren entweder auf statistischen Variabeln oder auf Zufallsvariablen, welche mittels Wahrscheinlichkeitsrechnungen bestimmt werden können. In vielen der aktuell verwendeten Verfahren werden mehrere dieser Variablen gleichzeitig untersucht. Dieses Vorgehen werden auch multivariate Verfahren genannt.

Häufig wird auch Machine Learning als Data Mining bezeichnet. Dies ist historisch aber nicht ganz korrekt. Während es beim Data Mining darum geht, neue Muster und somit Erkenntnisse aus Daten zu gewinnen, versucht man mit Machine Learning neue Berechnungsfunktionen aus den Daten abzuleiten. Moderne Data Mining Verfahren verwenden jedoch meist Techniken aus dem Bereich des Machine Learning. Im Folgenden werden Techniken und Verfahren gezeigt, die teilweise aus einer Überschneidung von Data Mining und Machine Learning bestehen. [2]

1.3 Beispiele von Anwendungsfälle

1.3.1 Watson for Oncology

Watson for Oncology (WFO) ist eine von IMB entwickelte Plattform, die bei der Behandlung von Krebspatienten helfen soll. Diese Plattform ist ein spannendes Beispiel dafür, wie Data Mining und Machine Learning den Menschen unterstützen kann, jedoch zeigt es auch, wo die Schwierigkeiten bei einem solchen System entstehen.

WFO wurde als Unterstützung für Ärzte*innen entwickelt, welche entscheiden, welche Krebsbehandlung eine krebskranke Person erhalten soll. Dabei kennt WFO eine vorgegebene Liste an Behandlungen. Aufgrund der Patientenakte entscheidet WFO anschliessend, welcher diese Behandlungen empfohlen werden können. WFO entwickelt also nicht eigenständig neue Behandlungen, sondern wählt aus bekannten Methoden die besten aus.

Diese Auswahl basiert auf den Informationen des Patientendossiers. Dies beinhaltet einerseits persönliche Merkmale wie das Alter oder das Geschlecht, aber auch die gesundheitlichen Informationen wie die Art und Stufe des Tumors, sowie die bereits durchgeführten Vorbehandlungen. Diese Daten werden mit früheren Patientendaten verglichen, bei welchen die damals gewählte Methodik für die Behandlung bekannt ist. Aufgrund dieses Vergleichs schlägt WFO die besten Methoden vor und schliesst auch gewisse Methoden aus.

Das behandelnden Ärzteteam nutzen anschliessend dieses Resultat als weitere Hilfestellung bei der Entscheidung, welche Behandlungsmethode gewählt wird. WFO trifft also nicht direkt eine Entscheidung, sondern dient lediglich als Hilfsmittel. Die effektive Behandlungsart wird aktuell immer von Ärzteteam ausgewählt.

Diverse Studien zu WFO haben ergeben, dass WFO als Hilfsmittel hilfreich sein kann, jedoch noch diverses Schwierigkeiten aufweist, bevor es die Evaluation eines Ärzteteams komplett ersetzen kann.

Das Hauptproblem dabei ist, dass WFO nicht vollständig transparent ist, aufgrund von welchen Informationen welche Behandlung vorgeschlagen wurde. Dies führt dazu, dass das Ärzteteam die Vorschläge nicht zwangsläufig nachvollziehen kann und allfällige Unstimmigkeiten im WFO-Algorithmus nicht erkannt werden können.

Ein weiteres Problem ist die Datengrundlage. WFO proklamiert eine Übereinstimmung von rund 93% zwischen den von WFO vorgeschlagenen Behandlungen und den von Fachpersonen vorgeschlagenen Behandlungen. Unabhängige Untersuchungen haben nun jedoch ergeben, dass dies stark davon abhängt, welche Art von Krebs vorliegt und in welcher Patientengruppe (Alter, Geschlecht) die zu behandelnde Person ist. [3] [4] [5]

1.3.2 Netflix Thumbnails

Sobald es um Gewinnung von Informationen aus Daten geht, wird oft Netflix genannt. Netflix hat sich in den letzten Jahren nicht zuletzt durch gesammelte Daten einen grossen Marktanteil im Online Streaming erarbeitet. Dabei trackt Netflix seine Nutzenden und versucht aufgrund des Verhaltens dieser Nutzenden herauszufinden, wieso eine Person ein Film oder eine Serie konsumiert.[6]

Anschliessend wird versucht, dieses Verhalten künstlich zu erzeugen. Das einfachste Beispiel dazu ist die Generierung der Thumbnails (Vorschaubilder). Je nach Präferenz der aktuell eingeloggten Person werden andere Bilder angezeigt. Im Rahmen dieser Arbeit wurde ein kleiner Selbsttest durchgeführt. Auf fünf verschiedenen Konten wurde nach der britischen Serie Peaky Blinders gesucht. Dabei wurden drei unterschiedliche Thumbnails angezeigt:

Abbildung 1: Angezeigte Thumbnails bei verschiedenene Netflix Konten (Quelle: Netflix)

Das linke Bild wurde nur in einem Konto angezeigt, während das mittlere und das rechte jeweils in zwei der fünf Konten angezeigt wurden. Spannend dabei ist das mittlere Bild. Eine der beiden Personen gab an, dass sie kürzlich eine Serie mit Anya Taylor-Joy (der Schauspielerin im Bild) geschaut hat. Die andere Person gab an, die Schauspielerin nicht zu kennen, nannte aber eine Vorliebe für Liebesdramen. Beides scheinen hier gute Gründe zu sein, dieses Thumbnail den anderen Thumbnails vorzuziehen. Die beiden Personen, welche das rechte Bild angezeigt erhielten, gaben an, allem voran Dramen und Thriller zu konsumieren, was das Thumbnail ebenfalls erklären könnte. Für das linke Bild konnte kein ausschlaggebendes Kriterium erkannt werden, da die Person keine spezifische Film-Vorliebe angeben konnte.

Obwohl hier über die Gründe nur gemutmasst werden kann, zeigt es doch, welche Kriterien Netflix für die Anzeige eines Thumbnails verwendet und wie dadurch die Aufmerksamkeit der Nutzenden geweckt werden soll.

1.3.3 Erkennung von Verbrechensmustern

Bereits 1949 veröffentlichte George Orwell das Buch mit dem Titel 1984. Dabei beschrieb er einen totalitären Überwachungsstaat, welcher fast die gesamte Privatsphäre seiner Bürger*innen unterdrückt. Dabei wird das Verhalten der einzelnen Personen überwacht und falsches, dem Staat nicht genehmes Verhalten bestraft. Dieser Ansatz wird heutzutage in gewissen Gebieten in die Wirklichkeit umgesetzt. Durch die Digitalisierung kann die Bevölkerung heute genauer untersucht und überwacht werden. Ein Beispiel dazu ist die Erkennung von Verbrechensmustern. Dabei wird versucht in bekannten Verbrechensdaten Muster zu erkennen, die Aufschluss darüber geben sollen, wann und wo das nächste Verbrechen verübt werden soll.

Ein gutes Beispiel dazu ist das Chicago Police Department. Dieses wendet Data Mining Techniken auf polizeiliche Datensätze an, darunter Kriminalitätsvorfälle, Verhaftungen und Wetterdaten. Diese historischen Daten werden mit IoT-Echtzeitdaten kombiniert (z. B. mit sensorgesteuerten Kameras). Dadurch können Verbrechen zeitnah oder sogar vor der Ausübung erkannt werden.

Durch dieses Verfahren können auch Regionen erkannt werden, in welchen überdurchschnittlich viele Verbrechen verübt werden. Diese Regionen können anschliessend genauer überwacht und geprüft werden. Wie Orwell bereits früh erkannt hat, führt dies auch zu diversen Nebeneffekten, die vielfach als Negativ aufgefasst werden. Auf einige davon wird im Kapitel Ethische Grundsätze eingegangen.[7]

Ein einfaches Verfahren zur Gruppierung von Verbrechen in Regionen, Zeiteinheiten oder Verbrechensart ist das Clustering. Dieses wird im Kapitel Clusteranalyse genauer erläutert. Dabei werden die bestehenden Verbrechen in Gruppen eingeteilt. Durch diese Gruppen können anschliessend Erkenntnisse gefunden werden, welche Regionen besonders gefährdet sind, oder welche Faktoren zu Verbrechen führen.[8]

Diese Art von Verbrechenserkennung hat durch die Digitalisierung stark zugenommen und es ist davon auszugehen, dass diese Analysen in Zukunft exakter und ausführlicher werden dürften.

1.3.4 Bayer

Ein kleines aber sehr spezifisches Beispiel für die Verwendung von Machine Learning und Data Mining ist die Unkrauterkennung der Firma Bayer.

Bayer hat über die Jahre rund 100'000 Bilder von Unkraut gesammelt. Daraus ergibt sich eine Datenbank von 70 Arten an Unkraut. Damit ein Bauer nun das Unkraut möglichst effektiv vernichten kann, nutze er diese Datenbank als Grundlage. Er kann ein Foto des Unkrautes hochladen und erhält die Information, um welches Unkraut es sich handelt.

Dadurch können die Landwirte die Auswirkungen ihrer Entscheidungen – zum Beispiel die Wahl des Saatguts, die Menge an Pflanzenschutzmitteln oder den Erntezeitpunkt – viel genauer vorhersagen.[9]

1.4 Ethische Grundsätze

Grundsätzlich gibt es viele ethische Überlegungen zu klären, sobald aus Daten Informationen gewonnen werden sollen. Diese alle zu erkennen und zu listen ist nahezu unmöglich. Deshalb werden im folgenden beispielhaft zwei Problematiken gezeigt, die bei der Erhebung, sowie der Verwendung der Daten auftreten. [10]

Da es sich hierbei um ethische Fragen handelt, können diese hier auch nicht abschliessend beantwortet werden.

1.4.1 Datenschutz / Privatsphäre

Zwei Themen die ethisch eine ähnliche Problematik darstellen sind der Datenschutz und die Privatsphäre der Menschen. Die Frage hierzu lautet: Wie viele Informationen über uns sind wir bereit bekannt zu geben, um die Vorteile von Datenanalysen zu erzeugen und wem vertrauen wir diese an?

Hierzu gibt es diverse rechtliche Grundlagen, die es zu befolgen gibt. Inwiefern diese eingehalten werden, ist jedoch für die meisten Menschen nur schwierig zu prüfen.

Das Thema Datenschutz wurde in den letzten Jahren zunehmend auch medial immer wieder behandelt. Zwei berühmte, eher negative behaftete Beispiele hierzu sind der von Edward Snowden publik gemacht Datenskandal rund um die NSA und der Datenskandal rund um Cambridge Analytica. In zweiterem Beispiel wurden Facebook Daten verwendet, um gezielt Wahlstimmen zu gewinnen. Ein guter Einstieg in die Thematik bietet der Film «The great hack» von Jehane Noujaim and Karim Amer oder die hier im Dokument verlinkten Quellen. [11] [12]

Bezüglich Privatsphäre können wir hier das Beispiel der Digidog aufführen. Hierbei handelt es sich um hundeähnlichen Roboter, welcher mit verschiedener Sensoren seine Umgebung überwacht. Das New York Police Department hat begonnen, diese als Hilfsmittel zu verwenden. Dies hat in der Bevölkerung von New York diverse Zweifel zur Privatsphäre verursacht.[13]

1.4.2 Bias

Unter einem Bias versteht sich im Allgemeinen eine Verzerrung der Wahrnehmung. Diesen gibt es nicht nur im Bereich Data Mining, sondern es handelt sich um ein generelles Problem im Umgang mit Informationen. Ein Beispiel dazu ist der Confirmation Bias. Dieser beschreibt, dass der Mensch gerne seine eigene Meinung bestätigt. Bei einer Recherche wird also nach Informationen gesucht die eine eigene Aussage belegen anstelle von anderen Quellen, die diese Aussage widerlegen könnten.[14]

Auch im Rahmen von Data Mining und Machine Learning existieren solche Bias. Zwei davon werden nachfolgend erläutert: [15]

1.4.2.1 Sample Bias (Stichprobenverzerrung)

Der Sample Bias tritt auf, wenn die Daten, die zum Trainieren eines Modells verwendet werden, nicht genau die Umgebung repräsentieren, in der das Modell arbeiten wird.

Als Beispiel können wir eine Software nehmen, die Tiere auf Bilder erkennen soll und so Hunde und Katzen unterscheiden kann. Zur Erstellung dieser Software verwenden wir nur Bilder, die im Innern aufgenommen wurden. Falls wir jetzt versuchen, mittels unserer Software Fotos zu klassifizieren, die im Wald aufgenommen wurden, werden wir mehr falsche Klassifizierungen erhalten als wir erwarten würden.

1.4.2.2 Prejudice Bias (Vorurteilsbedingte Verzerrung)

Wir wollen nun eine Software entwickeln, die uns im Bewerbungsprozess unterstützt und uns geeignete Dossiers vorselektiert. Für die Erstellung dieser Software verwenden wir die Daten aller aktuell in der Firma tätigen Personen. Da es sich in diesem Beispiel um eine IT-Firma handelt, besteht diese aktuell hauptsächlich aus Männern. Die Folge dieses Prejudice Bias ist nun, dass wir ausschliesslich Dossiers von Männern vorgeschlagen erhalten und keine von Frauen. Hätten wir in unseren Ursprungsdaten das Geschlecht nicht mit einbezogen und nur die Qualifikation der Personen angegeben, wäre dieser Bias zu verhindern gewesen.

2 Verfahren

Die Verfahren des Data Mining / Machine Learning können grob in zwei Teilgebiete unterteilt werden.[16][17]

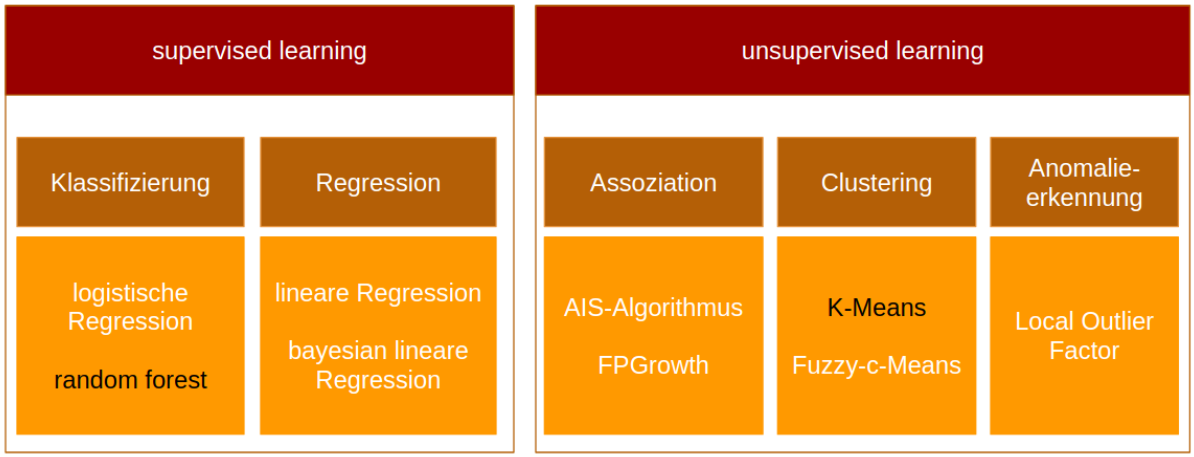


Abbildung 2: Übersicht der Verfahren (Quelle: Eigenkreation)

2.1 Supervised Learning

Beim supervised Learning (zu deutsch: überwachtes Lernen), sind die Zielwerte bekannt. Diese sind entweder aufgrund von Naturgesetzen oder Expertenwissen gegeben. Ziel dieser Vorgehen ist es, einen Eingabewert einer dieser bekannten Zielwerte zuzuweisen.

2.1.1 Klassifikationsverfahren

Abbildung 3: Klassifizierung von E-Mails (Quelle: Eigenkreation)

Bei der Klassifikation geht es darum, einen Datensatz einer Klasse aus einem gegeben Klassenpool zuzuweisen. Das einfachste Beispiel einer Klassifikation ist das Erkennen von Spam-Mail. Hierbei haben wir den Klassenpool (Spam, nicht-Spam). Ein eingehendes Mail wird nun entweder der Klasse Spam oder der Klasse Nicht-Spam zugewiesen. Der Algorithmus, welcher diese Zuteilung macht,

nennt sich Klassifikator. Beim Mailbeispiel sprechen wir von einer binären Klassifikation, da wir genau zwei Klassen in unserem Klassenpool haben. Demgegenüber steht die Multiklassenklassifizierung, bei der mehr als zwei Klassen im Klassenpool bestehen. Ein Beispiel hierzu ist die Klassifizierung von Früchten in die Klassen (Apfel, Birne, Orange, Aprikose). Zu beachten gibt es hierbei, dass mit der Anzahl Klassen auch die Schwierigkeit der korrekten Klassifikation anwächst. Deshalb ist häufig eine binäre Klassifikation einer Multiklassenklassifizierung vorzuziehen, sofern sich das Problem entsprechend vereinfachen lässt.

Beispiele von Klassifikationsverfahren sind die logistische Regression oder Random Forest.

2.1.2 Regressionsanalyse

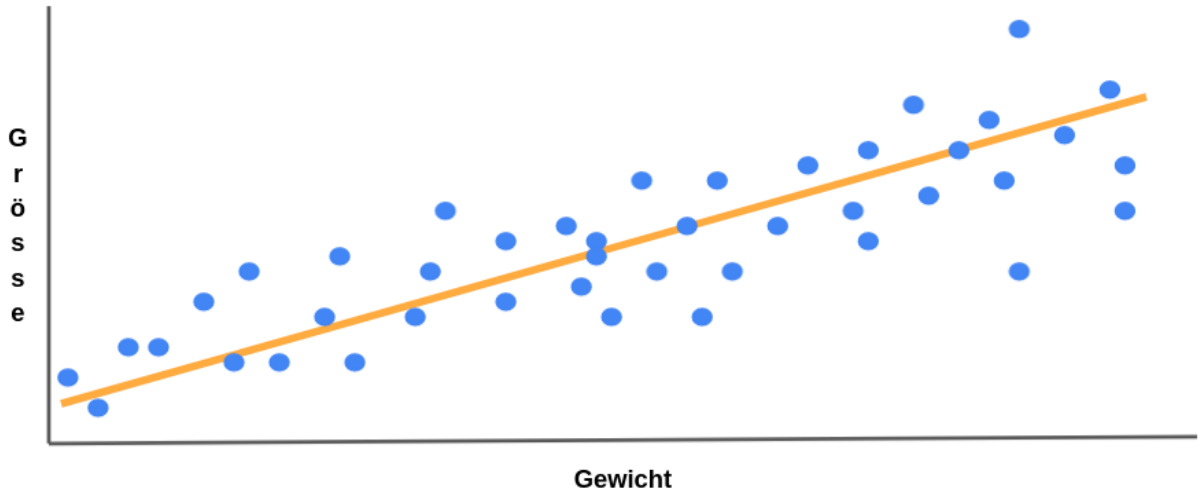


Abbildung 4: Bestimmung der Grösse durch Gewicht (Quelle: Eigenkreation)

Bei der Regressionsanalyse wird versucht, mittels einer abhängigen und einer oder mehreren unabhängigen Variablen ein Zusammenhang zu modellieren. Wichtig ist dabei die Abhängigkeit der ersten Variable. Fehlt diese funktioniert die Regression nicht. So ist es als Beispiel möglich, mittels Regression das Gewicht eines Menschen anhand seiner Körpergrösse zu modellieren, da grössere Menschen eher ein höheres Gewicht vorweisen. Jedoch ist es nicht möglich, aufgrund der Körpergrösse eines Menschen sein Jahreseinkommen vorauszusagen, da die beiden Variablen unabhängig voneinander sind. Eine einfache und häufige Art ist die lineare Regression. Hierbei wird versucht eine Funktion zu finden, die den Zusammenhang zwischen den Variablen möglichst genau beschreibt. Bei einer Einflüssgrösse (z. B. der Körpergrösse) und einer Zielgrösse (z. B. dem Gewicht), führt dies zu einer linearen Funktion. Diese Funktion kann auf verschiedene Arten berechnet werden. Eine gängige Möglichkeit bildet hier die Methode der kleinsten quadratischen Abweichung. Dabei wird die lineare Funktion gesucht, welche zu allen gegebenen Datenpunkten die kleinstmögliche quadratische Abweichung erzielt.

Ein weiteres Beispiel für die Regression ist die bayesian lineare Regression.

2.2 Unsupervised Learning

Beim unsupervised Learning (zu deutsch: unüberwachtes Lernen) sind im Gegensatz zum supervised Learning die Zielwerte nicht bekannt. Bei diesen Verfahren wird versucht innerhalb der Daten Muster zu erkennen, damit neue Erkenntnisse aus den Daten gezogen werden können.

2.2.1 Assoziationsanalyse

Kunde	Einkauf
1	{ Wasser, Windeln, Bier, Salz, Zopf, Gurken }
2	{ Betonmischer, Beton, Rose, Wein }
3	{ Kinderauto, Windeln, WC-Papier, Bier}
4	{ Chips, Sirup, Milch, Gurken, Brot, Senf }
5	{ Windeln, Marmelade, Essig, Bier}

Oft zusammen:
{ Windeln, Bier }

Assoziation:
{ Windeln } => { Bier }

Abbildung 5: Warenkorbanalyse (Quelle: Eigenkreation)

Bei der Assoziationsanalyse wird versucht eine Korrelation zwischen gemeinsam auftretenden Items herzustellen. Ein Item ist dabei ein Element einer gegebenen Menge. Durch Assoziation soll also herausgefunden werden, welche Items häufig gemeinsam auftreten. Ein häufiges Beispiel dazu ist die Analyse von Warenkörben. Dort wird versucht herauszufinden, welche Produkte häufig miteinander gekauft werden. So wurde als Beispiel herausgefunden, dass Männer zwischen 30 und 40 Jahre am Samstag häufig Windeln und Bier zusammen kaufen.

Zwei Beispiele für die Assoziationsanalyse sind der AIS-Algorithmus und FPGrowth.

2.2.2 Clusteranalyse

Bei der Clusteranalyse geht es darum Daten in verschiedene Gruppen einzuteilen. Das Ziel ist, dass jede Gruppe Elemente beinhaltet, die sich möglichst ähnlich sind. Inwiefern sich die Elemente ähnlich sind, spielt dabei für die Clusteranalyse keine Rolle. So kann die Ähnlichkeit bei Personendaten zum Beispiel aufgrund des Einkommens entstehen oder auch aufgrund der Herkunft der Personen. Auch eine Einteilung nach Alter wäre in so einem Falle denkbar. Wie die Ähnlichkeit bestimmt wird, hängt also davon ab, welche Daten (z. B. nur das Einkommen oder nur das Geschlecht) wir dem Algorithmus liefern.

Deshalb sind diese Gruppen am Anfang noch nicht zwangsläufig bekannt und werden erst durch die Analyse bestimmt. Sobald die Gruppen generiert wurden, kann untersucht werden welches die

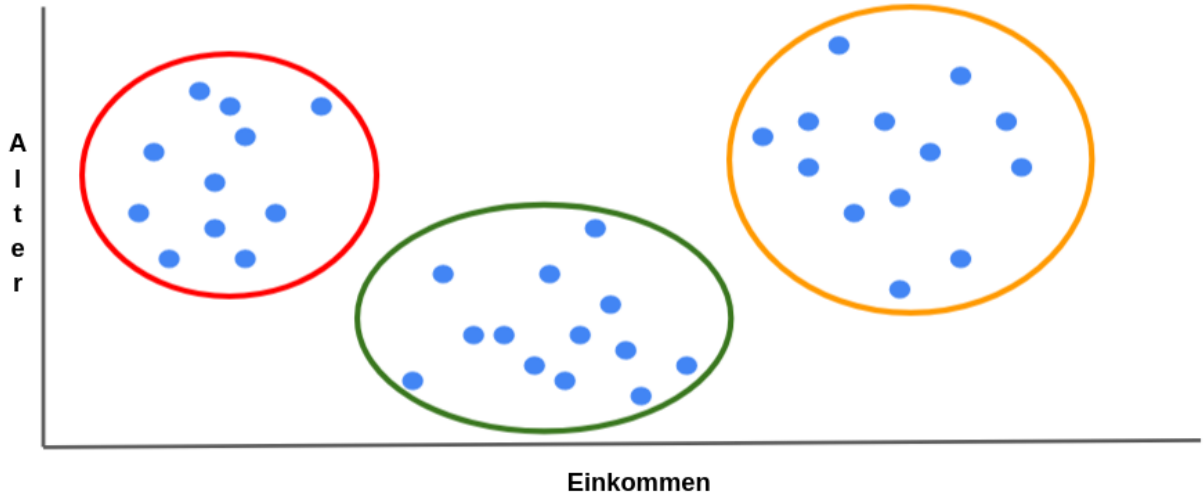


Abbildung 6: Gruppieren durch Alter und Einkommen (Quelle: Eigenkreation)

bestimmenden Merkmale einer Gruppe sind. Dies kann der Algorithmus nicht selber interpretieren. Die Interpretation der Gruppen ist ein separater, oft manueller Schritt im Prozess. Hier sind allem voran auch die Gruppen spannend, die wir nicht erwartet haben, da durch diese neue Informationen gewonnen werden.

Oft verwendete Verfahren sind hierbei der K-Means-Algorithmus und der Fuzzy-c-Means-Algorithmus.

2.2.3 Anomalieerkennung

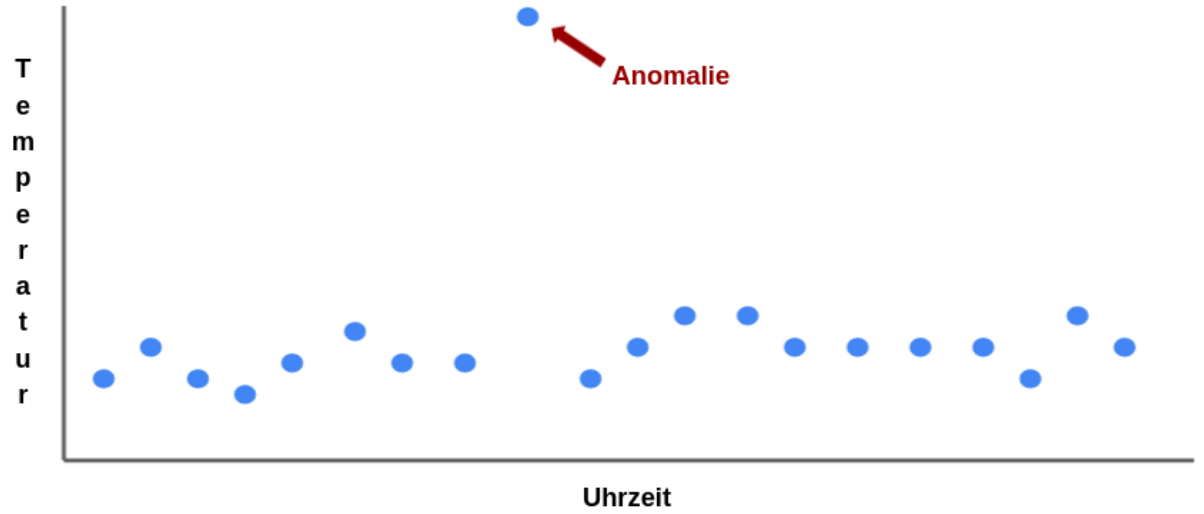


Abbildung 7: Falsche Werte erkennen bei Wetterstation (Quelle: Eigenkreation)

Die Anomalieerkennung wird verwendet, um zu erkennen, welche Datensätze nicht in ein bekanntes Muster passen. Es wird also konkret nach Ausreissern gesucht. Ein Beispiel hierzu sind Kreditkartendaten. Bei einer Kundin die normalerweise jeweils Beträge zwischen 50 und 200 Franken bezieht,

wäre ein Bezug von mehreren Tausend Franken ein Ausreisser. Hierbei könnte es sich um eine betrügerische Aktion handeln. So könnte in diesem Beispiel durch den Ausreisser der Diebstahl der Karte festgestellt werden. Ein weiteres Beispiel kann sein, Fehler in den Datenbeständen zu identifizieren. Eine Temperaturmessstation, die meist Werte zwischen 0 und 20 Grad liefert weisst wahrscheinlich eine Fehlfunktion auf, sollte sie einen Wert von 50 Grad liefern. Durch die Anomalieerkennung können solche Fehler ebenfalls schnell gefunden werden.

Ein bekannter Algorithmus für diese Erkennung ist der Local Outlier Factor.

3 Beispiele von Verfahren

3.1 K-means

K-means ist ein Clustering Algorithmus und somit ein unsupervised Verfahren. Das Ziel ist es Daten zu Gruppieren, die nicht vorher mit einem Label versehen wurden.

Der Trainingsschritt besteht daraus, die Trainingsdaten in K Gruppen oder eben Cluster zu verteilen. Anschliessend können wir dem trainierten Modell ein Datensatz füttern und erhalten als Resultat den zugehörigen Cluster.

Als Beispiel können wir hier wieder Netflix zur Hand nehmen. Netflix kann mittels K-means seine Nutzenden in eine beliebige Anzahl Cluster einteilen lassen. Dieser Trainingsschritt kann mittels den bestehenden Benutzerdaten gemacht werden. Meldet sich nun eine weitere Person an, kann diese aufgrund seiner Daten anschliessend einer dieser Gruppen zugewiesen werden. Somit kann Netflix dieser Person Vorschläge machen, die auf diesem Cluster basieren. Haben als Beispiel 80% der Nutzenden aus diesem Cluster den Film «the social dilemma» gesehen, schlagen wir diesen Film vor.

3.1.1 K in K-means

Das K in K-means gibt dabei vor, wie viele Cluster wir generieren wollen. Dieser Wert K wird durch die Person definiert, die den Algorithmus trainiert und nicht vom Algorithmus selber. Somit nennt sich K ein Hyperparameter. Hyperparameter sind alle Konfigurationen die vor dem Training durch den Menschen bestimmt werden und nicht durch den Algorithmus.

3.1.2 Funktionsweise

K-Means arbeitet mit sogenannten Zentroiden. Dabei gibt es pro Cluster einen Zentroiden. Dieser Zentroiden ist der Mittelpunkt des Clusters. Die Distanz zwischen zwei Punkten kann mittels der euklidischen Distanz berechnet werden. Der Zentroiden ist anschliessend der Wert mit den im Schnitt kleinsten euklidischen Abstand zu jedem Element im Cluster.

Der Ablauf von K-Means lässt sich also folgendermassen zusammenfassen:

1. Zuerst bestimmen wir K zufällige Zentroiden
2. Jeder Datenwert wird nun dem nächstliegenden Zentroiden zugewiesen. Dadurch werden alle Elemente eindeutig einem Cluster zugewiesen.
3. Nun berechnen wir pro Cluster den Zentroiden neu, damit dieser möglichst einen kleinen Abstand zu allen Elementen im Cluster hat.

4. Durch die Verschiebung der Zentroiden ist es nun möglich, dass ein Datensatz nicht mehr im Cluster mit dem nächstgelegenen Zentroiden ist. Deshalb führen wir die Schritte 2 und 3 solange aus, bis bei einem Durchgang kein Datensatz mehr den Cluster wechselt. Es können auch andere Abbruchbedingungen definiert werden, wie zum Beispiel eine maximale Anzahl an Durchgängen.

3.1.3 Bemerkungen

- K muss bereits vor Beginn des Trainings definiert werden. Diese Wahl ist somit essenziell für die Zugehörigkeit eines Elementes zu einer Gruppe. Es gibt verschiedene Techniken, die Qualität von den Clustern unter verschiedenen K zu prüfen. Zwei davon sind die Elbow-Methode und der Silhouettenkoeffizient.
- Ebenfalls relevant für die Qualität der Cluster ist die initiale Platzierung der Zentroiden. Je nach Platzierung können mit denselben Daten andere Cluster entstehen. Auch hier gibt es hilfreiche Algorithmen, wie k-means++. Bei k-means++ werden die Zentroiden so weit voneinander wie möglich platziert, was zu besseren Resultaten führt als die komplett zufällige Platzierung.

3.1.4 Notebook

Im Internet finden sich viele gute Python Notebooks zu K-means. Einen guten Überblick über K-means findet sich als Beispiel hier:

`https://colab.research.google.com/github/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/05.11-K-Means.ipynb`

3.1.5 Beispiel

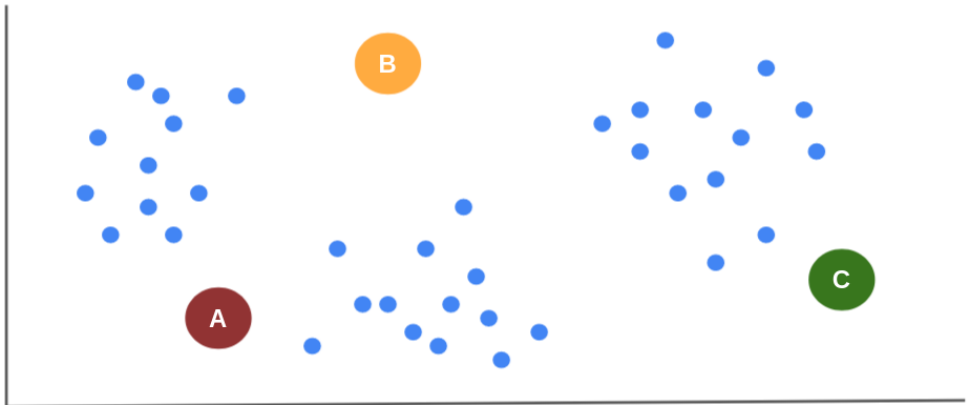


Abbildung 8: Initiale Platzierung der Zentroiden (Quelle: Eigenkreation)

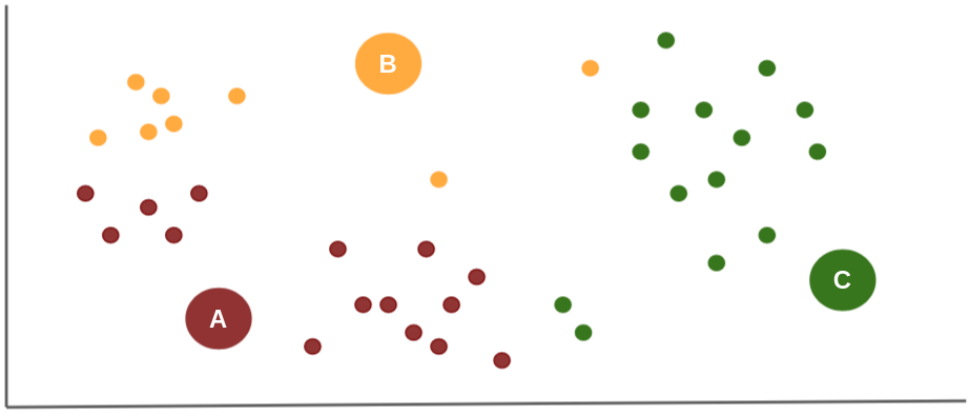


Abbildung 9: Zuweisen der Elemente zum nächsten Zentroiden (Quelle: Eigenkreation)

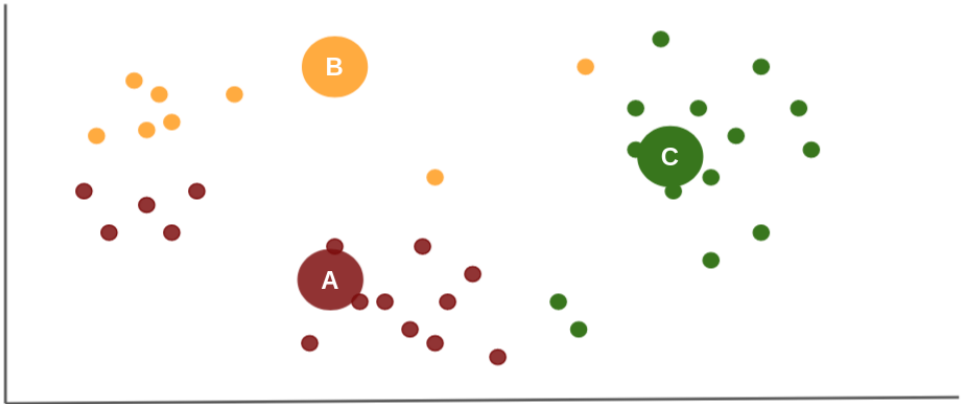


Abbildung 10: Neuplatzierung der Zentroiden (Quelle: Eigenkreation)

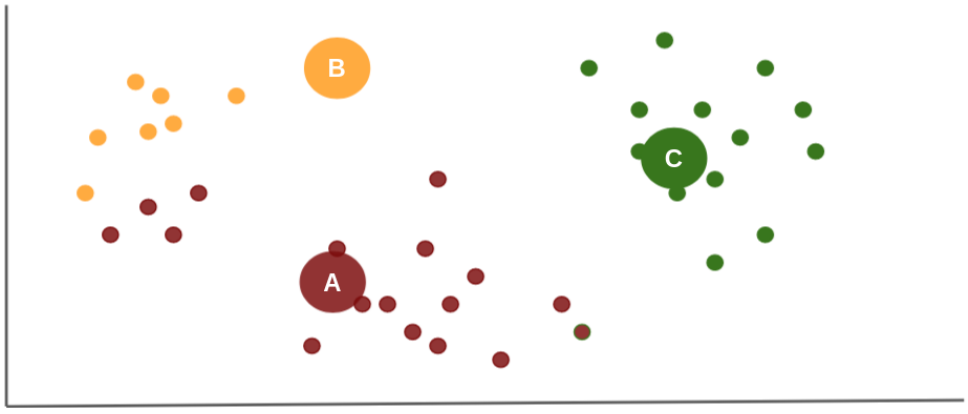


Abbildung 11: Neue Zuweisung der Elemente zum nächsten Zentroiden (Quelle: Eigenkreation)

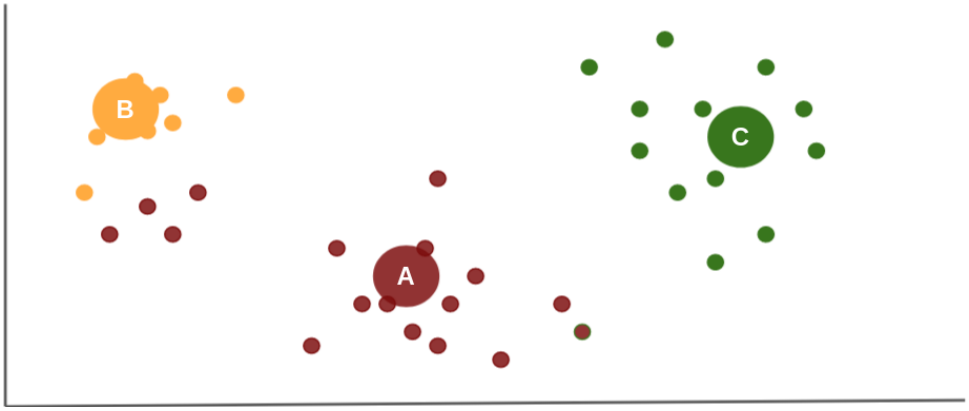


Abbildung 12: Erneute Neuplatzierung der Zentroiden (Quelle: Eigenkreation)

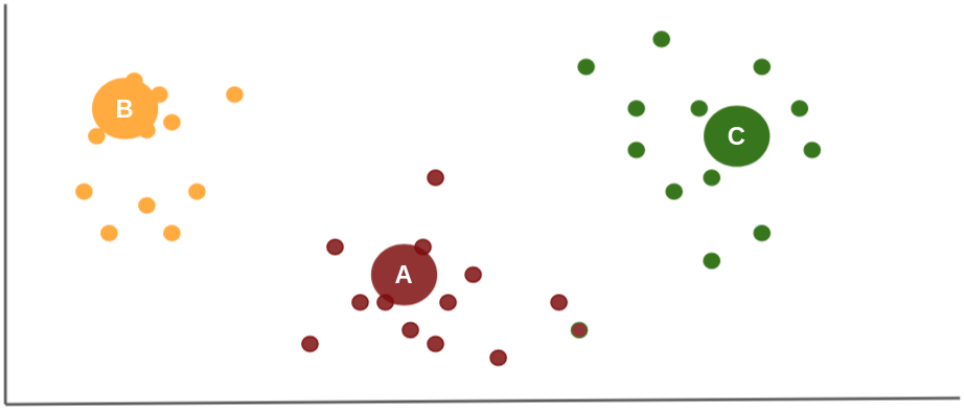


Abbildung 13: Neue Zuweisung der Elemente zum nächsten Zentroiden (Quelle: Eigenkreation)

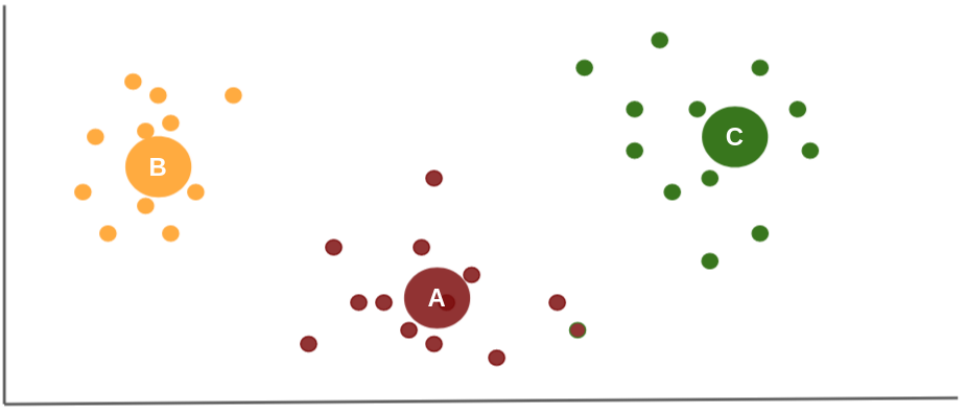


Abbildung 14: Sobald es keine Verschiebung mehr gibt, sind wir fertig (Quelle: Eigenkreation)

3.2 Random Forest

3.2.1 Decision Tree

Damit das Konzept von Random Forest erläutert werden kann, muss zuerst das Konzept eines Decision Tree (Entscheidungsbaum) erläutert werden. Ein Decision Tree kann für die Klassifizierung oder eine Regression verwendet werden. In unserem Beispiel verwenden wir den Decision Tree für die Klassifizierung von Datensätzen.

Sonnig	Wochenende	Temperatur > 30 Grad	Label
ja	ja	ja	ja
ja	nein	nein	ja
nein	ja	ja	ja
nein	nein	nein	nein
ja	ja	nein	nein
ja	nein	ja	ja
ja	ja	nein	nein
nein	ja	nein	nein

Abbildung 15: Trainingsdaten (Quelle: Eigenkreation)

In diesem Beispiel haben wir Datensätze, bei welchen jeweils angegeben ist, ob es an einem Tag sonnig war, ob es über 30 Grad warm war und ob es sich um einen Samstag oder Sonntag (Wochenende) gehandelt hat. In der Spalte <Label> ist angegeben, ob eine Person an diesem Tag in die Aare gesprungen ist oder nicht.

Aus diesen Daten wollen wir nun einen Decision Tree erstellen. Dafür wird nacheinander eines der Attribute (<Sonnig>, <Wochenende>, <Temperatur>) genommen und die Datensätze nach diesem Attribut getrennt. In welcher Reihenfolge wir die Attribute abarbeiten können wir mittels verschiedenen Methoden bestimmen. Das Ziel davon ist jeweils, nach der Aufteilung durch das Attribut möglichst alle Elemente mit <Label> ja auf einer Seite zu haben und alle anderen auf der anderen Seite. Zwei gängige Verfahren sind die Berechnung der Entropie oder des Gini-Faktors. In unserem Beispiel würde der Baum dann so aussehen:

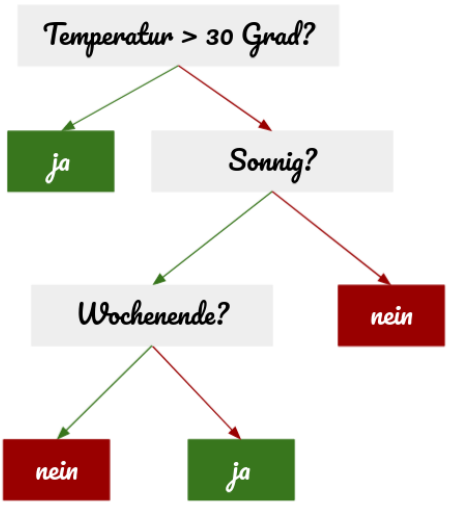


Abbildung 16: Trainingsdaten (Quelle: Eigenkreation)

Erläuterung zur Grafik: Hier wurde zuerst nach der Temperatur gefragt, da alle Datensätz mit einer Temperatur von über 30 Grad dem Label <Ja> zugeordnet werden können. Somit haben wir auf der linken Seite nur noch Datensätze mit dem Label <Ja>. Auf der rechten Seite fragen wir uns nun, ob es sonnig war. Bei allen nun übrig gebliebenen Datensätzen bei denen dies nicht zutrifft, lautet das Label <Nein>. Somit bleibt uns als letzte Frage das Wochenende. Bei allen übrig gebliebenen Datensätzen, an denen Wochenende war, lautet das Label <Nein>.

Das Problem bei solchen Decision Trees ist, dass sie sehr schnell overfitted sind. In unserem Beispiel sehen wir, dass alle Testdaten am Ende korrekt zugewiesen werden würden. Allerdings sehen wir auch, dass ein Eintrag, bei dem die Temperatur grösser 30 Grad war, immer zum Label <Ja> führt in unserem Baum. Diese Bedingung muss aber in zukünftigen Daten nicht zwangsläufig gegeben sein. Wir haben also einen Baum erstellt, der perfekt zu den Testdaten, nicht aber zwangsläufig zu weiteren Eingaben passt. Genau dieses Verhalten nennen wir Overfitting.

Anmerkung: Grundsätzlich muss ein Decision Tree nicht alle Testdaten korrekt klassifizieren (meist ist dies auch gar nicht möglich). Hier können wir z. B. angeben, dass wir auch mit 95% korrekten Klassifizierungen zufrieden sind.

3.2.2 Random Forest

Um das Problem des Overfitting zu beheben, verwenden wir nun anstelle eines Decision Tree einen Random Forest. Ein Random Forest besteht dabei aus mehreren Decision Trees. Dabei haben wir die Möglichkeit bei der Erstellung der einzelnen Trees jeweils nur eine gewisse Anzahl an Daten (Reihen oder Spalten) zu verwenden. Dadurch werden die einzelnen Bäume auf die gesamte Datenmenge diverser, was unser Overfitting verhindert.

Um anschliessend einen Datensatz zu klassifizieren, lassen wir diesen durch alle Trees in unserem

Forest laufen. Anschliessend wählen wir demokratisch das häufigste Resultat als korrekt.

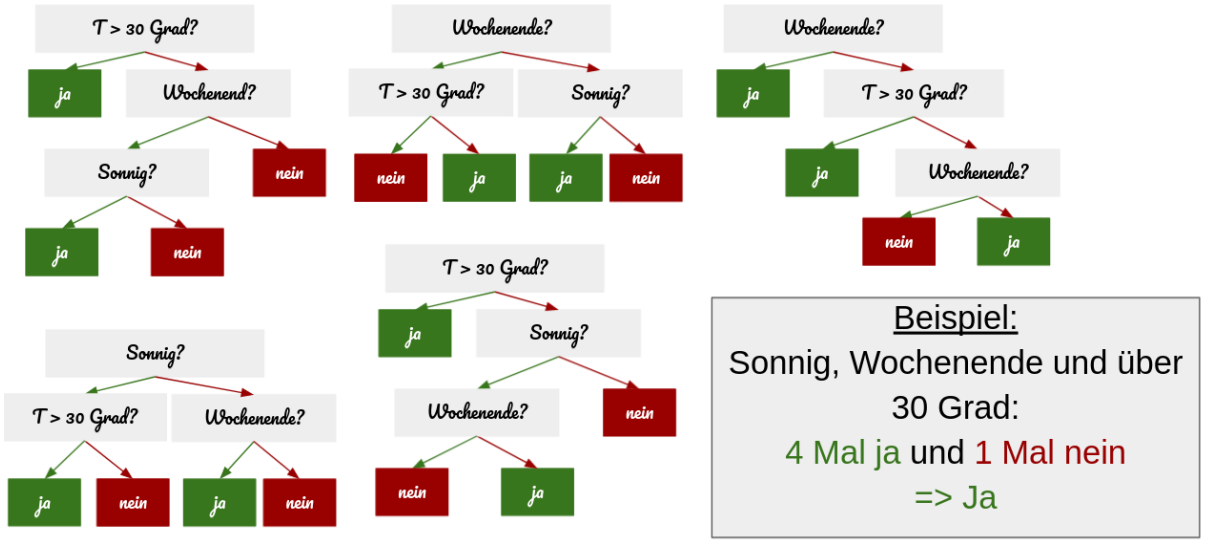


Abbildung 17: Random Forest (Quelle: Eigenkreation)

Erläuterung zur Grafik: Hier haben wir fünf verschiedene Trees erstellt, indem wir jeweils nur einen Bruchteil der Testdaten für das Training verwendet haben. Deshalb erhalten wir fünf verschiedene Trees. Um nun das korrekte Label zu bestimmen, prüfen wir unseren Datensatz mit allen Trees. In diesem Beispiel ergeben vier Trees ein ja und nur einer ein nein. Deshalb wählen wir als Label das Ja aus.

3.2.3 Notebook

Auch zu Random Forest finden sich viele Beispiele im Internet. Als Einstieg kann folgendes Notebook empfohlen werden: `https://github.com/WillKoehrsen/Machine-Learning-Projects/blob/master/random_forest_explained/Random%20Forest%20Explained.ipynb`

Literatur

[16] Sebastian Raschka und Vahid Mirjalili. *Python Machine Learning - Second Edition*. Birmingham, UK: Packt Publishing Ltd., 2017.

Online Quellen

[1] Laurenz Wuttke. *Data Mining: Algorithmen, Definition, Methoden und Anwendungsbeispiele*. https://datasolut.com/was-ist-data-mining/. [Online; accessed 22-Mai-2021].

[2] Johanna Ronsdorf. *Microsoft erklärt: Was ist Data Mining? Definition und Funktionen*. https://news.microsoft.com/de-de/microsoft-erklaert-was-ist-data-mining-definition-funktionen/. [Online; accessed 22-Mai-2021]. 2020.

[3] Virginia Lau. *How Watson for Oncology is advancing cancer care*. https://www.mmm-online.com/home/channel/technology/how-watson-for-oncology-is-advancing-cancer-care/. [Online; accessed 19-Mai-2021]. 2016.

[4] IBM. *Supporting cancer research and treatment*. https://www.ibm.com/watson-health/solutions/cancer-research-treatment. [Online; accessed 19-Mai-2021].

[5] Ezio Di Nucci Aaro Tupasela. *Concordance as evidence in the Watson for Oncology decision-support system*. https://link.springer.com/content/pdf/10.1007/s00146-020-00945-9.pdf. [Online; accessed 19-Mai-2021]. 2019.

[6] Allen Yu. *How Netflix Uses AI, Data Science, and Machine Learning — From A Product Perspective*. https://becominghuman.ai/how-netflix-uses-ai-and-machine-learning-a087614630fe. [Online; accessed 19-Mai-2021]. 2019.

[7] Jen Clark. *Facing the threat: Big Data and crime prevention*. https://www.ibm.com/blogs/internet-of-things/big-data-crime-prevention/. [Online; accessed 19-Mai-2021]. 2017.

[8] Shyam Varan Nath. *Crime Pattern Detection Using Data Mining*. http://cs.brown.edu/courses/csci2950-t/crime.pdf. [Online; accessed 19-Mai-2021].

[9] *Nachhaltige Lebensmittelproduktion durch gezielt ausgewählte Agrar-Daten*. https://www.talend.com/de/customers/bayer-digital-farming-gmbh/. [Online; accessed 19-Mai-2021].

[10] CHRIS BOUSQUET. *Mining Social Media Data for Policing, the Ethical Way*. https://datasmart.ash.harvard.edu/news/article/mining-social-media-data-policing-ethical-way. [Online; accessed 20-Mai-2021]. 2018.

[11] The Guardian. *Edward Snowden: the whistleblower behind the NSA surveillance revelations*. https://www.theguardian.com/world/2013/jun/09/edward-snowden-nsa-whistleblower-surveillance. [Online; accessed 20-Mai-2021]. 2013.

[12] The Guardian. *The Cambridge Analytica scandal changed the world – but it didn't change Facebook*. https://www.theguardian.com/technology/2019/mar/17/the-cambridge-

analytica-scandal-changed-the-world-but-it-didnt-change-facebook. [Online; accessed 20-Mai-2021]. 2019.

[13] Nur Ibrahim. *Does the NYPD Use Robot Dogs in Its Police Work?* https://www.snopes.com/fact-check/nypd-robot-dogs-police/. [Online; accessed 20-Mai-2021]. 2021.

[14] Anja Rassek. *Bias: Diese 7 kognitiven Verzerrungen sollten Sie kennen*. https://karrierebibel.de/bias/. [Online; accessed 20-Mai-2021]. 2019.

[15] Glen Ford. *4 human-caused biases we need to fix for machine learning*. https://thenextweb.com/news/4-human-caused-biases-machine-learning. [Online; accessed 20-Mai-2021]. 2018.

[17] guru99. *Data Mining Tutorial: What is | Process | Techniques and Examples*. https://www.guru99.com/data-mining-tutorial.html. [Online; accessed 22-Mai-2021].

Abbildungsverzeichnis