MASTERARBEIT

LERNFÄHIGKEIT VON DATA MINING MODELLEN ZUR AUTOMATISIERUNG IM BEREICH DES SMART HOMES

ausgeführt am

Studiengang
Informationstechnologien und Wirtschaftsinformatik

Von: Christian Edelsbrunner
Personenkennzeichen: 1610320015

Graz, am 04.12.2017 .....

Unterschrift

EHRENWÖRTLICHE ERKLÄRUNG

Ich erkläre ehrenwörtlich, dass ich die vorliegende Arbeit selbstständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen nicht benützt und die benutzten Quellen wörtlich zitiert, sowie inhaltlich entnommene Stellen als solche kenntlich gemacht habe.

.....

Unterschrift

DANKSAGUNG

Ich möchte mich an dieser Stelle bei all denen bedanken, die mich bei der Anfertigung meiner Masterarbeit so kräftig motiviert und unterstützt haben. Ganz besonders bedanken möchte ich mich bei meinem Betreuer Ing. Dipl.-Ing. Patrick Schweighofer BSc, für die sehr schnelle Reaktionszeit bei Fragen und für das immer konstruktive Feedback während der Erstellung meiner Masterarbeit. Sehr bedanken möchte ich mich auch bei meiner Frau und meiner kleinen Tochter, ohne deren Rücksichtnahme, Vertrauen und Unterstützung ich das berufsbegleitende Studium wohl nie bis zu dieser Arbeit gebracht hätte. Daneben gilt mein Dank Karl Jokesch und Julia Edelsbrunner, die in zahlreichen Stunden meine Arbeit Korrektur gelesen haben. Sie wiesen auf Schwächen hin und konnten als Fachfremde immer wieder zeigen, wo meine Formulierungen nicht ausreichend oder zu kompliziert waren. Nicht zuletzt gebührt meinen Eltern Dank, ohne welche ich wohl nie so weit gekommen wäre.

KURZFASSUNG

Die vorliegende Arbeit beschäftigt sich mit der Lernfähigkeit von Data Mining Modellen im Bereich des Smart Homes. Es wird der Frage nachgegangen, welche Auswirkungen ein kontinuierliches Training von Data Mining Modellen zur intelligenten Automatisierung der Beleuchtungssteuerung in einem Smart Home, auf die Fehlerrate des Data Mining Modells hat. Ziel ist es zu klären, ob durch kontinuierliches Training, auch zeitnah, die Fehlerrate von Data Mining Modellen zur Beleuchtungssteuerung im Smart Home verbessert werden kann.

Zu Beginn werden die Begriffe Data Mining und Machine Learning mit Hilfe einer Literaturrecherche, im Bereich des Smart Homes, definiert und ein Überblick über die Funktionsweise von Data Mining geschaffen, sowie das allgemeine Vorgehen zur Implementierung einer Data Mining Anwendung anhand des standardisierten CRISP-DM Prozesses erörtert. Anschließend wurden für die vorliegende Arbeit geeignete Data Mining Verfahren mit Hilfe von Literatur erörtert. Aus diesen Data Mining Verfahren und den erhobenen Anforderungen an den Anwendungsfall der vorliegenden Arbeit, wurde das Data Mining Verfahren der Entscheidungsbäume für die Umsetzung des Prototyps ausgewählt.

Im zweiten Teil der Arbeit wurde mit Hilfe des CRISP-DM Prozesses, ein Prototyp zur intelligenten Automatisierung der Beleuchtungssteuerung entwickelt und in ein bestehendes Smart Home integriert. Dieser Prototyp wurde im Zuge einer einfachen Fallstudie evaluiert. Dabei wurde die Fallstudie in zwei Iterationen evaluiert. In der ersten Iteration wurden einmalig trainierte Data Mining Modelle evaluiert und in der zweiten Iteration wurden kontinuierlich trainierte Data Mining Modelle evaluiert. Basierend auf der Auswertung der Daten und dem Vergleich der Vorhersageleistung, konnte die Erkenntnis gewonnen werden, dass ein kontinuierliches Training zeitnah zu einer Verbesserung der Vorhersageleistung beitragen kann.

ABSTRACT

This thesis investigates the learning ability of data mining models in the area of smart homes. Data mining is well understood, but the learning ability of data mining models in smart homes has so far been neglected. The paper begins with a literature review to define the terms ‘data mining‘ and ‘machine learning‘ in the domain of smart homes. These definitions are built on by a literature research to clarify how data mining functions and how it can be implemented using a standardized process. Afterwards, this theoretical information is used for an advanced literature research about data mining methods which match the requirements for the application in this thesis. This section concludes with the data mining method decision trees chosen for the implementation of a prototype. The second part of the paper describes the development of a prototype of a prediction and decision-making system for light automation in a smart home. This prototype was used to evaluate the learning ability of data mining models in the area of smart homes. A single case study is chosen and analyzed with two iterations. The first iteration evaluates the prediction performance of data mining models trained once. The second iteration evaluated the performance of continuously trained data mining models. Based on the comparison of both iterations, it is shown that continuously trained data mining models in this domain have far fewer error rates than data mining models trained once.

INHALTSVERZEICHNIS

1 EINLEITUNG

Die Forschung im Themengebiet Smart Home ist seit mehr als einem Jahrzehnt sehr aktiv und dieser Trend nimmt weiter zu (Priti & Yatin, 2016). Der Schwerpunkt der Entwicklungen in diesem Bereich lag in den letzten Jahren auf rein technologischen Themen wie Automatisierung und Vernetzung der beteiligten Systeme. Wie am Beispiel von Pan und Liang (2017) zu sehen ist, bilden die aus diesen Forschungen entstandenen Technologien, Methoden und Erkenntnisse, die Ausgangsbasis für Forschungen im Bereich von intelligenteren Systemen, welche als digitale und selbst lernende Systeme das tägliche Leben im Smart Home erleichtern können. Augusto und Nugent (2006) argumentieren, dass die Funktionalität von Smart Homes über Automatisierung und Vernetzung hinausgehen sollte und sich Smart Homes durch den Einsatz von selbstlernenden Algorithmen zu intelligenteren Gebäuden entwickeln können. Diese Gebäude können, ohne explizites Regelwerk, für Vorgänge im Smart Home, Aktionen vorhersagen und Entscheidungen treffen, womit sie auf intelligente Art und Weise das Leben der Bewohner und Bewohnerinnen erleichtern.

In der Literatur finden sich bereits einige Arbeiten, welche sich mit dieser Entwicklung des Smart Homes und konkreter Umsetzungen beschäftigen. So haben zum Beispiel Dixit und Naik (2014) gezeigt, dass es möglich ist, Verhaltensmuster in einem Smart Home anhand von gesammelten Sensordaten, zu erkennen. Als weiteres Beispiel können Das, Chen, Seelye und Cook (2011) genannt werden, sie haben gezeigt, dass Vorhersagealgorithmen in einem Smart Home für die Erkennung von komplexen Abläufen im Smart Home verwendet werden können und das sich daraus Systeme erstellen lassen, welche es ermöglichen Aufgaben in einem Smart Home zu automatisieren.

In der vorliegenden Arbeit soll die Fähigkeit zur kontinuierlichen Verbesserung und Optimierung der Vorhersagequalität von solchen Systemen, aus den Bereichen Data Mining und Machine Learning, für den Einsatz in intelligenten Vorhersage- und Entscheidungssystemen, im Bereich des Smart Homes, evaluiert werden. Dabei soll untersucht werden, ob durch ein kontinuierliches Aktualisieren der Entscheidungsbasis für ein solches System, bessere Ergebnisse erzielt werden können, als bei einem System, welches einmalig mit einem definierten Datenbestand erstellt und anschließend nicht mehr aktualisiert wird. Durch diese Evaluierung sollen Erkenntnisse darüber gewonnen werden, ob es ausreichend ist Data Mining Modelle für dieses Anwendungsgebiet einmalig zu trainieren oder ob der höhere Aufwand, welche durch die Implementierung eines kontinuierlichen Trainings entsteht, durch die bessere Lernfähigkeit, aufgewogen wird oder gar notwendig ist um die Vorhersagequalität zu stabilisieren.

1.1 Ausgangssituation

In der Literatur finden sich bereits einige Publikationen, die sich mit dieser Art der Weiterentwicklung im Bereich des Smart Homes beschäftigen. Viele dieser Arbeiten beschäftigten sich mit der Erweiterung von Smart Homes, im Sinne von Wohneinheiten mit vernetzten Sensoren und Aktoren, um intelligente Vorhersage- und Entscheidungssysteme. Einige Arbeiten fokussieren sich ausschließlich auf die Erkennung von Aktivitäten, Trends oder Ereignissen im Smart Home. Andere Arbeiten gehen bereits weiter und beschäftigen sich auch mit der Umsetzung von konkreten Assistenzsystemen oder Automatisierungslösungen, welche den Bewohnern und Bewohnerinnen das Leben erleichtern, indem sie Aufgaben übernehmen. Beispiele für solche Arbeiten sind die bereits genannten Arbeiten von Dixit und Naik (2014), Cook (2012) oder Pan, Liang und Xu (2017), sowie die, in weiterer Folge in der vorliegenden Arbeit genannten Arbeiten von Cook, Crandall, Thomas und Krishnan (2013), Cook et al. (2003), Tapia, Intille und Larson (2004), Youngblood, Heiermann, Holder und Cook (2015) oder Spring, Cook, Weeks, Dahmen und La Fleur (2017). Einige dieser Arbeiten entstammen dem MavHome[1] Smart Home Projekt der Universität Texas in Arlington. Ein weiterer Fokus der bestehenden Arbeiten liegt auf Anwendungen im Gesundheitsbereich, wie zum Beispiel die intelligente Überwachung von Alzheimerpatienten oder Alzheimerpatientinnen.

In der Literatur finden sich aber nur sehr wenige Publikationen, welche sich damit beschäftigen ob solche Lösungen, wenn sie einmal entwickelt und installiert wurden, auch so zuverlässig weiter funktionieren, wenn sich die Umwelt im Smart Home verändert. Es gibt sehr wenig Publikationen, welche sich mit der Leistung solcher Vorhersage- und Entscheidungssysteme im Zusammenhang mit sich ändernden Nutzungsverhalten im Smart Home, beschäftigen.

Die vorliegende Arbeit soll genau in diesem Bereich neue Erkenntnisse bringen. Mit der vorliegenden Arbeit sollen Informationen darüber gewonnen werden, wie sich solche Vorhersage- und Entscheidungssysteme im Bereich des Smart Homes im Einsatz verhalten. Es soll ermittelt werden, ob die Systeme, wenn sie einmal erstellt und installiert wurden, ihre Arbeit auch langfristig mit der ursprünglichen Qualität erfüllen können, oder ob eine kontinuierliche Anpassung der Systeme erforderlich ist, oder zumindest Vorteile bringen kann.

1.2 Ziele und Forschungsfrage

Das Ziel dieser Arbeit ist es, die Fähigkeit zur kontinuierlichen Verbesserung der Vorhersagequalität, eines auf Data Mining und Machine Learning Techniken basierendes Vorhersage- und Entscheidungssystems für die intelligente Beleuchtungssteuerung, in einem Smart Home zu evaluieren. Es wird die Fähigkeit der eingesetzten Techniken untersucht, sich durch kontinuierliches Training durch neue Ereignisdaten laufend zu optimieren und somit die Zuverlässigkeit der Vorhersagen zu verbessern und sich so auch automatisch an geändertes Benutzungsverhalten anzupassen. Kontinuierlich trainierte Data Mining Modelle, sind in der

[1] http://ailab.wsu.edu/mavhome/

Umsetzung und im Betrieb aufwändiger, als Modelle, welche nur initial vor der Inbetriebnahme mit Trainingsdaten auf ein bestimmtes Qualitätsniveau hin trainiert werden. Ein kontinuierliches Training bringt aber den Vorteil, dass das Modell sich stetig verbessern und sich so auch an geändertes Bewohnerverhalten anpassen kann. Mit Hilfe der vorliegenden Arbeit, soll eine Aussage getroffen werden können, ob kontinuierlich trainierte Data Mining Modelle für diesen Anwendungsfall gut geeignet sind und der Mehraufwand, den sie mit sich bringen, damit gerechtfertigt werden kann. Um dieses Ziel erreichen zu können, wird in der vorliegenden Masterarbeit, die folgende Forschungsfrage beantwortet:

Welche Auswirkung hat das kontinuierliche Training des Data Mining Modells für ein Vorhersagesystem zur intelligenten Automatisierung der Beleuchtungssteuerung in einem Smart Home, auf die Fehlerrate des Data Mining Modells?

Mit dieser Forschungsfrage und den erstellten Hypothesen soll die Frage beantwortet werden, ob ein laufendes Training des Data Mining Modells zur laufenden und zeitnahen Verbesserung der Fehlerraten beitragen kann oder ob auf das laufende Training verzichtet werden kann, da die, mit einem einmaligen Training, erhaltenen Vorhersagewerte nicht signifikant verbessert werden. So soll die Auswahl von Data Mining Techniken für diese Art von Vorhersagesystemen erleichtert werden, in dem auf Basis des Ergebnisses abgewogen werden kann, ob der höhere Implementierungsaufwand im jeweiligen Anwendungsfall, durch die Vorteile aufgewogen wird.

Für die vorliegende Arbeit wurden die folgenden Hypothesen aufgestellt:

Hypothese:

Ein laufendes Training des Data Mining Modells, für ein Vorhersagesystem zur intelligenten Automatisierung der Beleuchtungssteuerung, in einem Smart Home, hat zeitnahe Auswirkungen auf die Vorhersagequalität des Modells.

H1:

Ein laufendes Training des Data Mining Modells, für ein Vorhersagesystem zur intelligenten Automatisierung der Beleuchtungssteuerung in einem Smart Home, führt zur zeitnahen Verbesserung der Fehlerraten des Data Mining Modells.

H0:

Ein laufendes Training des Data Mining Modells, in einem Vorhersagesystem zur intelligenten Automatisierung der Beleuchtungssteuerung in einem Smart Home, führt zu keiner zeitnahen Verbesserung der Fehlerraten des Data Mining Modells.

Mit zeitnah ist in der vorliegenden Arbeit ein Evaluierungszeitraum von ein bis zwei Monaten definiert worden. So soll eine, für Bewohner recht zeitnahe Verbesserung des Data Mining Modells evaluiert werden.

1.3 Vorgehen und Methodik

Mittels einer Literaturrecherche wurden zuerst die grundlegenden Begriffe und die grundlegende Funktionsweise zu den Themen Data Mining, Machine Learning und Smart Home für die vorliegende Arbeit erörtert und definiert. Teil dieser Literaturrecherche war es auch in Fachliteratur, mit Fokus auf Fachbücher, aus dem Bereich Data Mining, nach standardisierten Prozessen zur Umsetzung von Data Mining Aufgaben zu suchen und einen der gefundenen Prozesse auszuwählen, welcher anschließend als Grundlage für die Umsetzung des Prototyps verwendet wurde. Aufbauend auf den Definitionen der Grundbegriffe und den damit verbunden Einschränkungen und Abgrenzungen, wurde in der Fachliteratur, mit Fokus auf Fachbücher, aber auch in wissenschaftlichen Zeitschriften und Konferenzprotokollen, nach passenden Data Mining Verfahren gesucht, die Funktionsweise dieser erörtert und ein Vergleich der gefundenen Data Mining Verfahren aufgestellt. Auf Basis dieses Vergleiches und den Anforderungen aus dem Bereich des Smart Homes, wurden abschließend Data Mining Verfahren für die vorliegende Arbeit und für die Umsetzung des Prototyps ausgewählt.

Für die Literaturrecherche nach Zeitschriftenartikeln und Konferenzprotokollen wurde primär auf die Online-Portale Scopus[2], Google Scholar[3] und Web-of-Science[4] zurückgegriffen, für die Suche nach Fachbüchern wird auf das Online-Portal Google Scholar und auf die Online-Bibliothekssuche der TU Graz, sowie der FH CAMPUS 02 zurückgegriffen.

Zur Evaluierung der Anpassungs- und Lernfähigkeit eines Data Mining Systems zur intelligenten Automatisierung der Beleuchtung in einem Smart Home, wurde die Durchführung und Auswertung einer Fallstudie als primäre Evaluierungsmethode ausgewählt. Die Fallstudie wurde als einfache Fallstudie in einem ausgewählten Smart Home über einen definierten Zeitraum und unter realen Bedingungen durchgeführt. Für die Durchführung der Fallstudie wurde ein, auf die zuvor ausgewählten Data Mining Verfahren basierendes System mit prototypischem Funktionsumfang entwickelt und in das Smart Home zur Evaluierung integriert.

Dieser Prototyp wurde auf Basis der Ergebnisse, der zuvor durchgeführten Literaturrecherche, dem ausgewählten Data Mining Prozess, sowie den ausgewählten Data Mining Verfahren aufgebaut und entwickelt. Die Basis für den Prototyp stellte die Anwendung und Implementierung des ausgewählten Data Mining Prozesses und der ausgewählten Data Mining Verfahren dar. Der Data Mining Prozess wurde angewendet und der daraus entstandene Prototyp des Vorhersage- und Entscheidungssystems wurde in das bestehende Smart Home integriert. Nach der Integration wurde ein Test des Prototyps durchgeführt, um die Funktionsfähigkeit für die Durchführung der Fallstudie zu prüfen.

Der Funktionsumfang des Prototyps beschränkt sich auf die Vorhersage des Ein- und Ausschaltens von Beleuchtungen. Das System unterstützt keine Vorhersage von Farbeinstellungen oder Helligkeitsstufen. Damit handelt es sich um ein klassisches

[2] http://www.scopus.com/

[3] https://scholar.google.at/

[4] https://apps.webofknowledge.com

Klassifizierungsproblem im Kontext von Data Mining und um keine Vorhersage von Trends oder Zahlenwerten.

Um die Ergebnisse der Fallstudie zu vervollständigen und methodische Schwächen zu kompensieren, wurde während der Fallstudie von den Teilnehmern und Teilnehmerinnen ein Tagebuch über Anomalien und unerwünschte Abweichungen in der intelligenten Beleuchtungssteuerung geführt, welches anschließend ausgewertet und im Ergebnis berücksichtigt wurde. Ergänzend wurde mit einem abschließenden Interview der Teilnehmer und Teilnehmerinnen, der subjektive Eindruck zur intelligenten Beleuchtungssteuerung und Qualität der Vorhersagen im Durchführungszeitraum ermittelt.

Die Datenerhebung zur Evaluierung des Prototyps und damit für die Beantwortung der Forschungsfrage, erfolgt zum einen durch laufende Auswertung der Fehlerraten des Data Mining Modells während der Evaluierung mit Testdaten und zum anderen durch die manuelle Datenerhebung in den Tagebuchprotokollen. Es ist notwendig die laufende Auswertung der Fehlerraten mit den Informationen aus den Tagebüchern zu kombinieren, um Fehler, welche vom Data Mining Modell als korrekt erachtet wurden, aber aus Sicht der Teilnehmer und Teilnehmerinnen unerwartet oder falsch waren, zu erkennen und berücksichtigen zu können. Zusätzlich zur Betrachtung der Entwicklung der Fehlerraten des Data Mining Modells und der Fehlereinträge in den Tagebuchprotokollen, wurden die im Durchführungszeitraum neu hinzugekommenen Ereignisse im Smart Home statistisch analysiert und mit den ursprünglichen Trainingsdaten verglichen, um Korrelationen oder Veränderungen im Nutzungsverhalten erkennen und in der Auswertung berücksichtigen zu können.

Die Fallstudie, sowie die Tagebuchführung und das abschließende Interview, wurden auf Basis der Vorgehensweise von bereits durchgeführten ähnlichen Evaluierungen aufgebaut und geplant. Für diesen Zweck wurde eine eigene Literaturrecherche durchgeführt, welche zum einen mit Fokus auf Fachbücher, die grundsätzliche Vorgehensweise einer solchen Evaluierung geklärt hat und zum anderen, mit Fokus auf wissenschaftliche Zeitschriften, nach ähnlichen Evaluierungen gesucht hat, welche als Grundlage für die Evaluierung in der vorliegenden Arbeit verwendet wurden. Diese Literaturrecherche wurde in denselben Portalen wie die vorhergehende Literaturrecherche durchgeführt.

1.4 Aufbau der Arbeit

Im ersten Inhaltskapitel der vorliegenden Arbeit werden die Grundbegriffe rund um die Themen Data Mining und Smart Home geklärt und definiert. Anschließend wird in diesem Kapitel auf die Grundfunktionalität von Data Mining und einen standardisierten Data Mining Prozess, sowie auf die dafür benötigten Techniken und Werkzeuge zur Umsetzung von Data Mining Projekten, eingegangen. In diesem Kapitel werden zusätzlich Einschränkungen und Abgrenzungen von Techniken aus dem Bereich von Data Mining für die vorliegende Arbeit getroffen, um die im zweiten Inhaltskapitel folgende Aufbereitung und Auswahl von Data Mining Verfahren gezielter auf die vorliegende Arbeit ausrichten zu können.

Im zweiten Inhaltskapitel wird auf die unterschiedlichen Gruppen von Data Mining Verfahren eingegangen, welche zur Eingrenzung der vorliegenden Arbeit passen und die wichtigsten Verfahren innerhalb dieser Gruppen werden aufbereitet. Auf Basis dieser Auflistung wird anschließend eine Auswahl von einer oder mehreren Data Mining Verfahren, für die Umsetzung des Prototyps in der vorliegenden Arbeit getroffen.

Im darauf folgenden Kapitel, wird auf Basis der definierten Grundlagen, dem standardisiertem Data Mining Modell und den gewählten Data Mining Verfahren ein Data Mining Modell für die vorliegende Arbeit aufgebaut und anhand des standardisierten Data Mining Prozesses durchgeführt und dokumentiert. Anschließend wird in diesem Kapitel auch auf den Entwurf und die Implementierung des Prototyps eingegangen, in welchem das zuvor entworfene Data Mining Modell integriert wird. Dieser Prototyp wird anschließend im bestehenden Smart Home, welches für die anschließende Evaluierung verwendet wird, integriert. In diesem Kapitel werden auch auf Technologie- und Architekturentscheidungen im Zuge der Prototypenimplementierung, sowie auf den genauen Funktionsumfang und die Funktionsweise des Prototyps eingegangen und entsprechende Auswahlen getroffen und begründet.

Im daran anschließenden Kapitel wird auf das Vorgehen zur Evaluierung des Prototyps und zur Durchführung der Fallstudie, sowie das abschließende Interview eingegangen. Es wird auf die Vorgehensweise für die Durchführung der Fallstudie, die begleitende Tagebuchführung und das abschließende Interview eingegangen. Abschließend wird die Durchführung der Evaluierung in diesem Kapitel geplant und dokumentiert.

Im Anschluss an die Evaluierung werden die Ergebnisse dieser gesammelt, ausgewertet und interpretiert. In diesem Kapitel werden die Daten für die Beantwortung der Forschungsfrage ausgewertet, Analysen zur Beantwortung der Forschungsfrage und die Überprüfung der Hypothesen durchgeführt.

Im abschließenden Kapitel werden zuerst die gewählten Definitionen, die Vorgehensweise zur Modellierung des Data Mining Modells, die Implementierung des Prototyps, sowie die Planung und Durchführung der Evaluierung kurz zusammengefasst und reflektiert. Anschließend werden die Ergebnisse und die Beantwortung der Forschungsfrage kurz zusammengefasst und reflektiert. Abschließend sind das Conclusio, sowie der Nutzen und Ausblick dokumentiert. In Abbildung 1 ist der Aufbau der vorliegenden Arbeit zur besseren Übersicht, grafisch dargestellt.

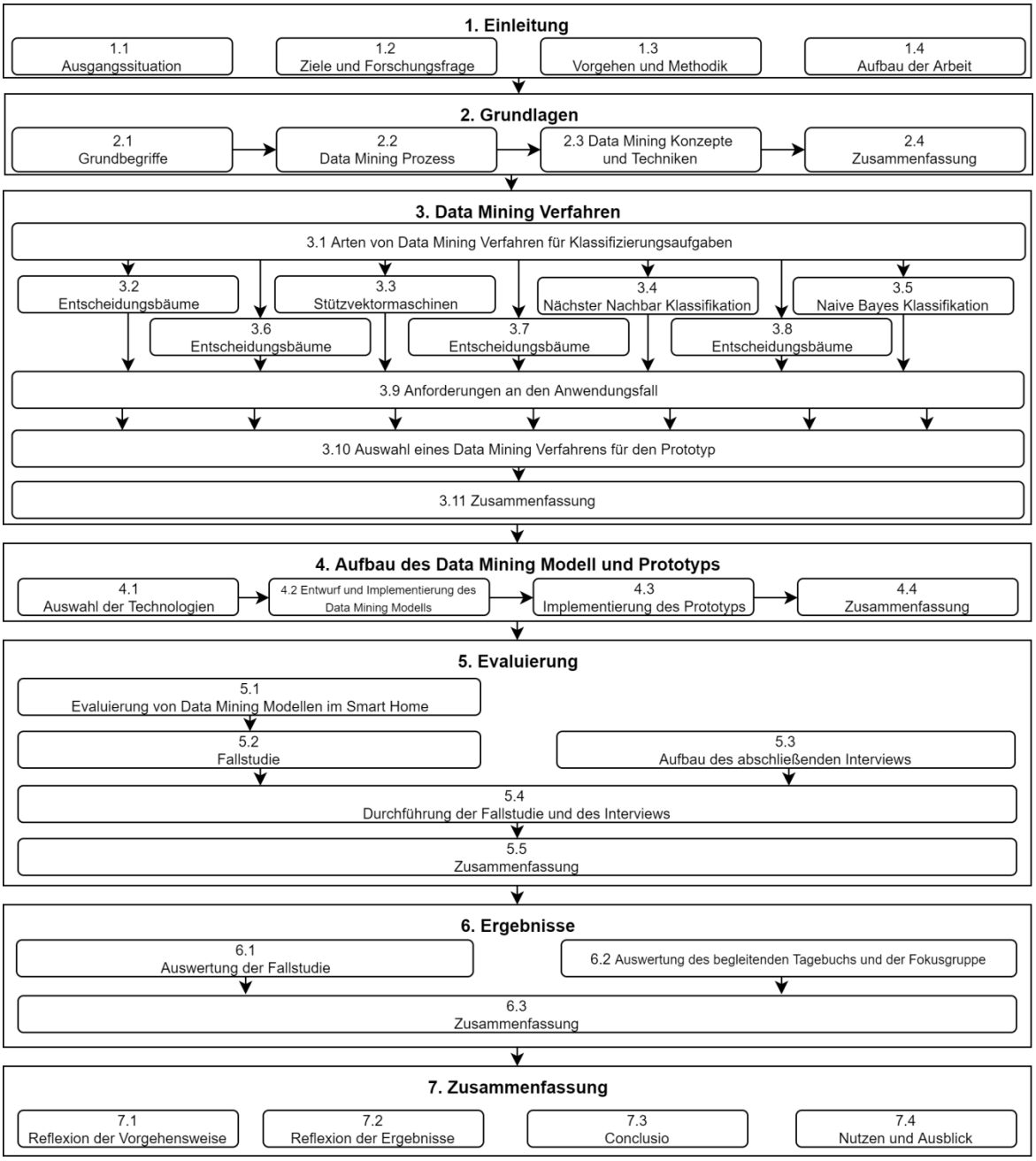

Abbildung 1: Aufbau der Arbeit

2 GRUNDLAGEN

In diesem Kapitel werden zuerst die Grundbegriffe rund um die Themen Data Mining, Machine Learning und Smart Home definiert. Anschließend wird die grundsätzliche Idee von Data Mining und das Vorgehen zur Anwendung dieser Idee, anhand eines standardisierten Data Mining Prozess beschrieben. Die Beschreibung dieses Prozesses wird abschließend, durch für die vorliegende Arbeit relevante Themen rund um Data Mining Modelle, deren Validierung und Verbesserung, ergänzt. Auf Basis des vorgestellten Prozesses wird in den folgenden Kapiteln der vorliegenden Arbeit, ein Prototyp aufgebaut. Als Ergebnis soll dieses Kapitel einen Überblick über die Funktionsweise von Data Mining und den wichtigsten Aspekten, welche für die Auswahl eines Data Mining Verfahrens in der vorliegenden Arbeit relevant sind, liefern.

2.1 Grundbegriffe

In diesem Abschnitt werden die Grundbegriffe Data Mining, Machine Learning und Smart Home für die vorliegende Arbeit mit Hilfe von Literaturrecherchen erörtert und definiert. Diese Definitionen stellen die Grundlage für Einschränkungen und Abgrenzungen in der vorliegenden Arbeit dar.

2.1.1 Smart Home

Es gibt unterschiedliche Definitionen des Begriffes Smart Home. In der bestehenden Literatur lassen sich zwei Definitionen des Begriffes ableiten, welche für den Kontext der vorliegenden Arbeit relevant sind. Schiefer (2015) reflektierte mehrere Definitionen des Begriffes mit Fokus auf Vernetzung und Kommunikation von Sensoren. In diesen Definitionen beschreibt sich ein Smart Home durch den Einsatz und der Vernetzung von intelligenten Sensoren und Aktoren in modernen Wohnungen und Häusern mit dem Ziel, Bereiche wie Licht, Temperatur oder Heizung über einen Computer oder ein Smartphone zentral steuern und mit definierten Regeln automatisieren zu können. Augusto und Nugent (2006) haben erkannt, dass der Begriff des Smart Homes weiter gehen sollte als nur Vernetzung und klassische regelbasierte Automatisierung. Ein Smart Home sollte mit Hilfe der vernetzten Sensoren und Aktoren in der Lage sein eine größere Intelligenz zu bilden. So soll ein Smart Home entsprechend dieser Definition in der Lage sein, aktuelle Ereignisse im Haus autonom zu erkennen und basierend darauf Vorhersagen, sowie Entscheidungen zu treffen und autonom in das Geschehen im Smart Home einzugreifen und Aktionen zu tätigen oder zumindest vorzuschlagen.

Smart Homes nach der ersten Definition, sind bereits als Produkte am Markt erhältlich, diese Produkte konzentrieren sich auf den Einsatz und die Vernetzung von intelligenten Sensoren und Aktoren zur einfachen Automatisierung und zentralen Steuerung von digitalen

Komponenten in modernen Wohnungen und Häusern. In solchen Smart Homes ist es einfach möglich Sensordaten zentral aufzuzeichnen, darzustellen und Aktoren zentral über das Smartphone zu steuern (Schulze-Sturm, 2016). Für die eigentliche Automatisierung, abseits der zentralen Steuerung von Komponenten, ist in diesen Produkten eine manuelle Konfiguration und Erstellung von Automatisierungsregeln notwendig welches sich in den aktuellen Produkten oft als aufwändiges Unterfangen herausgestellt hat (Jakobi, Ogonowski, Castello, Stevens, & Wulf, 2016).

Smart Homes nach der zweiten Definition sind noch nicht in vollen Umfang am Markt in Form von Produkten verbreitet. Der größte Unterschied zwischen Smart Homes der vorhergehenden und dieser Definition liegt in der Automatisierung von Aktionen und der Art und Weise wie Aktionen angestoßen werden. In Smart Homes entsprechend der vorhergehenden Definitionen werden Regeln für die Automatisierung von Aktionen explizit für jeden Fall manuell durch einen Menschen konfiguriert oder programmiert. Smart Homes mit größerer Intelligenz werden ergänzend zu den Komponenten der vorhergehenden Definition mit Algorithmen der künstlichen Intelligenz und des Data Mining ausgestattet, um auf Basis der vorliegenden historischen Informationen autonom Wissen über aktuelle und zukünftige Ereignisse zu generieren. So können solche Systeme lernen gewisse Aktivitäten in einem Smart Home zu identifizieren und Vorhersagen auf zukünftige Aktivitäten daraus abzuleiten. (Cook D. J., 2012)

Zusammengefasst lässt sich sagen, dass die erste Definition des Begriffes Smart Homes den Einsatz und die Vernetzung von intelligenten Sensoren und Aktoren zur zentralen Steuerung von Wohnungen und Häusern in Kombination mit klassischer Automatisierung beschreibt. Die zweite Definition von Smart Homes ergänzt die erste Definition um den Einsatz von künstlicher Intelligenz zur autonomen Automatisierung und Unterstützung. In der vorliegenden Arbeit wird ein Smart Home, in welchem bereits vernetzte Sensoren und Aktoren zur Automatisierung im Einsatz sind, um die Fähigkeit erweitert autonom auf Basis von Erfahrungswerten Vorhersagen zu treffen und aktiv in das Geschehen im Smart Home einzugreifen. Die Ausgangsbasis für die Evaluierung in der vorliegenden Arbeit bildet somit ein Smart Home der ersten Definition. Dieses wird im Bereich der Beleuchtungssteuerung durch die Entwicklung eines Systems zur Vorhersage von Ereignissen zu einem Smart Home nach der zweiten Definition welches aktiv und ohne explizite Programmierung eines Regelwerkes auf Basis von historischen Daten in die Steuerung des Smart Homes eingreift.

2.1.2 Data Mining und Data Mining Verfahren

Witten, Frank, Hall und Pal (2016) definieren Data Mining als Vorgang zur Mustererkennung aus großen Datenmengen. Sie beschreiben, dass der Vorgang an sich nichts Neues ist, seit Jahrhunderten versuchen Menschen, Muster in Daten zu erkennen und zu interpretieren, um neues Wissen über die Daten zu generieren. Der Begriff Data Mining selbst wurde in den letzten Jahrzehnten durch die immer größer werdenden digitalen Datenmengen, den dadurch entstandenen Auswertungsmöglichkeiten und den damit verbundenen technischen Herausforderungen geprägt. Han, Kamber und Pei (2012) definieren Data Mining im engeren Sinne als Schritt in einem übergeordneten Prozess der Wissensgenerierung aus einer großen

Datenbasis. In diesem Schritt erfolgt, wie in der Definition nach Witten, Frank, Hall und Pal (2016), die eigentliche Mustererkennung.

Dieser Data Mining Prozess umfasst aber auch weitere Schritte wie das Verständnis der Daten, die Auswahl und Vorbereitung dieser, sowie die Evaluierung des Data Mining Ergebnisses und die Präsentation des generierten Wissens. Dieser Prozess wurde sowohl in den Publikationen von Witten, Frank, Hall und Pal (2016), als auch von Han, Kamber und Pei (2012) beschrieben und definiert. Erstere definieren den gesamten Prozess als Data Mining Prozess, Letztere definieren diesen Prozess als Wissensgenerierungsprozess aus großen Datenbeständen und nur den eigentlichen Mustererkennungsschritt als Data Mining. Bereits Fayyad, Piatetsky-Sharpio und Smyth (1996) haben die Definitionen wie Han, Kamber und Pei (2012) beschrieben und die Wissensgenerierung aus großen Datenbeständen als den übergeordneten Prozess und Data Mining als Anwendung eines Verfahrens zur Mustererkennung definiert. In allen drei Publikationen wird der Prozess als iterativer Vorgang beschrieben in dem das Ergebnis iterativ evaluiert und optimiert werden muss. Han, Kamber und Pei (2012) kommen zum Schluss, dass ihre Definition zwar präziser ist, dass aber sowohl in der wissenschaftlichen, als auch in der industriellen Praxis der Begriff der Wissensgenerierung aus großen Datenbeständen kaum verwendet wird und in den Publikationen der letzten Zeit stattdessen meist vom Data Mining Prozess gesprochen wird. Sie verwenden diese Definition daher auch in Ihrer weiteren Arbeit. In dieser Definition ist Data Mining als Prozess zu sehen, in dem im zentralen Data Mining Schritt ein, oder mehrere, Data Mining Verfahren zur Mustererkennung und Analyse der Daten angewendet wird. Die Definitionen in allen drei untersuchten Publikationen decken sich in diesem Verständnis, nur die Bezeichnungen für den Prozess und den eigentlichen Schritt der Wissensgenerierung, weichen voneinander ab. Für die vorliegende Arbeit wird, wie in der Arbeit von Witten, Frank, Hall und Pal (2016) definiert und wie in der Arbeit von Han, Kamber und Pei (2012) geschlussfolgert, der Prozess als Data Mining Prozess betrachtet und der eigentliche Schritt des Data Mining als Anwendung eines Data Mining Verfahrens gesehen.

Durch die Anwendung von Data Mining Verfahren werden interessante Muster in großen Datenmengen gefunden und damit in weiterer Folge Wissen über die Daten generiert. Im Data Mining Prozess werden eine, oder auch mehrere dieser Methoden angewandt, um sogenannte Data Mining Modelle aus den Daten zu generieren. Diese Modelle repräsentieren die gefundenen Muster in den Daten und können anschließend genutzt werden um Wissen über die analysierten Daten zu generieren und Vorhersagen zu neuen Daten oder Trends für die zukünftige Entwicklung eines Wertes erzeugen (Witten, Frank, Hall, & Pal, 2016).

Han, Kamber und Pei (2012) listen grundsätzliche Data Mining Verfahren auf, mit welchen nach unterschiedlichen Arten von Mustern gesucht werden kann. Sie unterscheiden grundsätzlich zwischen den zwei Gruppen: Den beschreibenden und den voraussagenden Data Mining Verfahren. Beschreibende Data Mining Verfahren beschreiben Eigenschaften von Daten in der analysierten Datenbasis. Voraussagende Data Mining Verfahren verfolgen das Ziel durch Induktion aus den analysierten Daten, Voraussagen für neue Daten zu treffen. Für die vorliegende Arbeit liegt der Fokus auf voraussagende Data Mining Verfahren, da diese Art von Data Mining Verfahren sowohl für die Forschungsfrage, als auch die Hypothese relevant ist.

Zusätzlich wird zwischen Vorhersagesystemen unterschieden, mit welchen Klassifizierungen oder Trends und laufende Werte, wie exakte Zahlenwerte, vorhergesagt werden können (Witten, Frank, Hall, & Pal, 2016). Für die vorliegende Arbeit wird die Auswahl auf Data Mining Verfahren eingeschränkt welche sich für die Vorhersage von Klassifizierungen eignen, konkret werden die beiden Klassifizierungen, Beleuchtung ein und Beleuchtung aus, vorhergesagt.

Kotu und Deshpande (2015) nennen die Disziplinen der Statistik, der künstlichen Intelligenz und des Machine Learning als Kerndisziplinen für Data Mining Verfahren. Viele der heute existierenden Data Mining Verfahren sind aus einer, oder aus Kombinationen dieser Disziplinen entstanden. Witten, Frank, Hall und Pal (2016) kommen zum Schluss, dass man keine klare Grenze zwischen Machine Learning und Statistik im Kontext von Data Mining ziehen kann. Laut ihnen stammen viele Ansätze für die Erkennung von Mustern in großen Datenmengen aus der klassischen Statistik. In beiden Bereichen haben sich in der Domäne von Data Mining in den letzten Jahrzenten ähnliche Ideen und Techniken entwickelt. In vielen Machine Learning Techniken wurden daher Ansätze und Konzepte aus der Statistik integriert und spezialisiert. In der vorliegenden Arbeit wird dieses Themengebiet als Machine Learning definiert und rein statistische Methoden nicht getrennt behandelt. Mitchel (1997) ist ebenso zu diesem Schluss gekommen, zusätzlich sieht er im Bereich des Machine Learning auch die Implementierung und Verwendung vieler Konzepte aus dem Bereich der künstlichen Intelligenz, welche für die Anwendung von Machine Learning wiederverwendet und optimiert wurden. Somit sind laut diesen Autoren viele Data Mining Verfahren im Bereich des Machine Learnings zu finden. Hier finden sich auch einige Ansätze wieder, die eigentlich aus dem Bereich der Statistik, oder der künstlichen Intelligenz stammen und für Machine Learning wiederverwendet und optimiert wurden.

Für die vorliegende Arbeit lässt sich Data Mining somit als Prozess zur Wissensgenerierung aus einer großen Datenmenge durch Analyse dieser auf das Vorhandensein von interessanten Mustern definieren. Als zentraler Schritt dieses Prozesses werden Data Mining Verfahren auf den Datenbestand angewandt, welche die eigentliche Erkennung von interessanten Mustern durchführen. Der Prozess umfasst auch umliegende Schritte, wie die Vorbereitung und die Evaluierung der angewandten Verfahren, auf diese Schritte wird im folgenden Kapitel näher eingegangen. Die gefundenen Muster werden als Data Mining Modell bezeichnet und ermöglichen die Generierung von Wissen über die analysierten Datenbestände. Es gibt grundsätzlich zwei Gruppen von Data Mining Verfahren, die vorhersagenden und die beschreibenden Verfahren. Erstgenannte sind für die Beantwortung der Forschungsfrage relevant und werden in weiterer Folge in der vorliegenden Arbeit behandelt. Data Mining Verfahren entspringen den Gebieten der Statistik, der künstlichen Intelligenz und dem Machine Learning. In den Machine Learning Ansätzen finden sich viele Ideen und Methoden aus den beiden anderen Bereichen wieder, welche für den Einsatz im Gebiet des Data Minings optimiert wurden. In der vorliegenden Arbeit wird somit der Fokus auf Methoden aus dem Bereich des Machine Learnings gelegt, auf Machine Learning selbst wird im folgenden Abschnitt genauer eingegangen.

2.1.3 Machine Learning

Bereits Frawley, Piatetsky-Shapiro und Matheus (1992) haben den Trend und die Möglichkeiten gesehen Machine Learning Techniken für die Analyse von großen Datenmengen zur Wissensgenerierung anzuwenden und dafür als Überbegriffe Data Mining und Wissensgenerierung definiert. Fayyad, Piatetsky-Sharpio und Smyth (1996), Mitchell (1997), Han, Kamber, Pei (2012), Sugiyama (2015), als auch Witten, Frank und Pal (2016) sehen Methoden und Techniken aus dem Bereich des Machine Learning als die wichtigste Familie der Data Mining Verfahren. Nicht zuletzt, weil klassische Methoden aus den Bereichen der Statistik und der künstlichen Intelligenz in viele Machine Learning Ansätzen eingeflossen sind und für diesen Anwendungsbereich optimiert wurden. Im Kontext von Data Mining werden in allen der sechs zuvor genannten Publikationen, unter dem Begriff Machine Learning Methoden zusammengefasst, welche es ermöglichen, komplexe Muster in Daten zu erkennen und dabei die Fähigkeit des Lernens auf Basis der analysierten Daten aufbauen.

Han, Kamber und Pei (2012) fassen die Fähigkeit des Lernens durch Machine Learning Methoden für den Einsatz in Data Mining zusammen. Lernen in diesem Kontext beschäftigt sich damit, wie Programme komplexe Muster automatisiert erkennen können und auf Basis von Daten besser in diesem Vorgang werden und so die Leistungsfähigkeit der Erkennung steigern können. Witten, Frank, Hall und Pal (2016) kommen zu einem ähnlichen Schluss. Sie kommen aber auch zum Schluss, dass der Begriff Lernen in diesem Kontext nur sehr schwer zu definieren ist, da Lernen in der Theorie auch die Begriffe Denken und Zweck einschließt. Diese beiden Begriffe sind hier nur sehr schwer greifbar. Sie argumentieren, dass Machine Learning in diesem Anwendungsgebiet eher mit dem Begriff trainieren definiert werden kann, diese Definition hat aber auch ein paar Schwächen. Sie definieren Lernen im Kontext von Data Mining als eine Anwendung von praktischem Lernen. Konkret geht es darum, Muster in Daten zu erkennen und auf Basis dieser Wissen zu generieren oder Vorhersagen zu treffen. Dabei soll die Leistungsfähigkeit und die Präzision durch die Analyse weiterer Daten verbessert werden.

Han, Kamber, Pei (2012) nennen vier klassische Probleme aus dem Bereich des Machine Learnings, welche für Data Mining von großer Relevanz sind:

- Beaufsichtigtes Lernen: Im Kontext von Data Mining ist damit Lernen gemeint, bei dem eine Zielrichtung vorgegeben wird. Ein Anwendungsfall ist hier die Klassifizierung von Daten. Hier wird dem Algorithmus vorgegeben, in welche Klassen er die Daten einordnen soll.
- Unbeaufsichtigtes Lernen: Im Kontext von Data Mining ist damit Lernen gemeint, bei dem keine Zielrichtung vorgegeben wird. Ein Anwendungsfall ist hier das Clustering von Daten. Hier wird dem Algorithmus keine Vorgabe gemacht, es wird versucht ohne weitere Vorgaben Cluster in den Daten, auf Basis von Ähnlichkeiten oder Abhängigkeiten zu bilden.
- Kombination aus den beiden zuvor genannten Varianten: Durch die Kombination beider Varianten, kann zusätzliches Wissen generiert werden. Vor allem können Ausreißer in den Daten durch die Kombination besser berücksichtigt werden.

- Aktives Lernen: Hier wird der Anwender der Software aktiv gebeten, durch Interaktion Entscheidungen vorzugeben, wenn der Algorithmus selbst nicht sicher eine Entscheidung treffen kann.

Für die vorliegende Arbeit kommt primär das beaufsichtige Lernen in Frage. Für den Prototypen wird eine binäre Klassifizierung in Form einer Ja/Nein Antwort, auf die Frage „Soll das Licht X Ein oder Aus geschaltet werden?“, zu treffen sein. Es ist zu prüfen, ob durch eine Kombination einer Machine Learning Methode aus dieser Gruppe mit anderen Methoden ein besseres Ergebnis erzielt werden kann. Darauf wie eine solche Kombination mehrerer Data Mining Verfahren aussehen kann, wird in einem späteren Abschnitt dieses Kapitels eingegangen.

Zusammengefasst lassen sich unter dem Begriff Machine Learning im Kontext von Data Mining Technologien und Methoden zusammenfassen, die es ermöglichen komplexe Muster in Daten zu erkennen und auf Basis dieser Daten Wissen zu generieren oder Vorhersagen zu neuen Daten zu treffen. Dabei können diese Techniken durch neue Daten trainiert werden, um die Leistungsfähigkeit der Mustererkennung und die Fähigkeit zum Treffen von Vorhersagen, aus diesen Mustern, zu verbessern. Für die vorliegende Arbeit wird primär eine Methode aus dem Bereich des beaufsichtigten Lernens eingesetzt.

2.2 Data Mining Prozess

Chapman et al. (2000) veröffentlichten ein Whitepaper zu einem standardisierten und branchenübergreifenden Data Mining Prozess, welcher in Zusammenarbeit von mehreren großen Firmen, im Zeitraum zwischen 1996 und 2000 entwickelt wurde. Das Ziel der Entwicklung dieses Prozesses war es, Erfahrungen, bewehrtes Vorgehen, sowie bewehrte Techniken in einem offenen und freien Prozess zusammenzufassen. Auf diese Weise soll es Unternehmen einfacher möglich sein, mit Data Mining zu starten und dabei nach einem standardisierten und bewährten Prozess vorzugehen. Dieser Prozess heißt Cross Industry Standard Process for Data Mining oder kurz CRISP-DM. Er beschreibt grundlegende Prozessschritte für die Durchführung von Data Mining Aufgaben. Shearer (2000) hat dieses Whitepaper in einer wissenschaftlichen Publikation zusammengefasst und die Relevanz eines standardisierten Prozesses für Data Mining hervorgehoben und mit Beispielen ergänzt. Shearer (2000) ist dabei zum Schluss gekommen, dass der Prozess in der Wirtschaft angenommen wird und er es vor allem weniger erfahrenen Datenanalysten ermöglicht, schneller in die Materie einzusteigen und erfolgreiche Data Mining Projekte auf Basis dieses Prozesses umzusetzen.

In der Literatur finden sich auch andere Data Mining Prozesse wie der ursprüngliche Knowledge Discovery Process von Fayyad, Piatetsky-Shapior und Smyth (1996), welcher in vielen Teilen in CRISP-DM abgedeckt ist, oder dem Sample, Explore, Modify and Asses Prozess, kurz SEMMA, welcher von einem der größten Statistiksoftwareentwicklungsunternehmen entwickelt wurde. Neben diesen drei großen Prozessen gibt es noch eine Menge weiterer, im wesentlich weniger bekannte und oft stark spezialisierte Prozesse, welche zum Großteil die Konzepte aus diesen drei Prozessen aufgegriffen und weiterentwickelt oder adaptiert haben (Marbán, Mariscal, & Segovia, 2009). Marbán, Mariscal und Segovia (2009) haben CRISP-DM als de facto Standard

für Data Mining Projekte bezeichnet. Auch die Arbeit von Witten, Frank, Hall und Pal (2016) verwenden diesen Prozess als Data Mining Prozess. Azevedo und Santos (2008) haben die drei zuvor genannten großen Data Mining Prozesse verglichen und dabei festgestellt, dass CRISP-DM der Prozess ist, der die Anforderung der Prozesse in Summe am besten abdeckt. Aus den hier genannten Gründen wird der CRISP-DM für die vorliegende Arbeit als Vorlage für den zu implementierenden Data Mining Prozess gewählt und in Folge detaillierter beschrieben.

Der CRISP-DM Prozess, wie in Abbildung 2 dargestellt, beinhaltet die Schritte: Fachliches Verständnis, Verständnis der Daten, Vorbereitung der Daten, Modellierung, Evaluierung und Verwendung. Diese Schritte sind in einem iterativen Prozess angeordnet, welcher ein iteratives Durchlaufen aller Schritte vorsieht und so eine ständige Optimierung ermöglicht (Shearer, 2000).

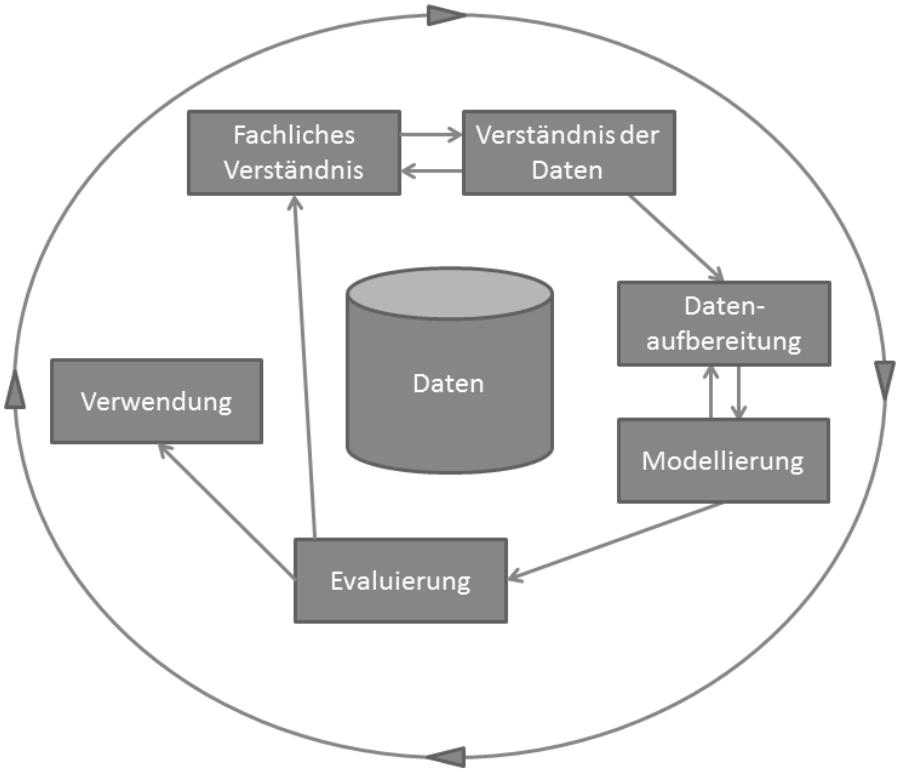

Abbildung 2: CRISP-DM Prozess. (Chapman, et al., 2000)

Der erste Schritt des Prozesses ist das fachliche Verständnis (Business Understanding). Dieser Schritt verfolgt das Ziel, das Data Mining Projekt aus fachlicher Sicht zu verstehen und eine Definition für das Data Mining Problem abzuleiten. Es soll ein Verständnis geschaffen werden, welche Daten analysiert werden müssen und welche Fragen beantwortet werden sollen (Chapman, et al., 2000). Witten, Frank, Hall und Pal (2016) sehen in diesem Schritt die Anforderungsanalyse als Hauptaufgabe. Welche Anforderungen werden aus fachlicher Perspektive an das Data Mining Modell gestellt und aus welchen Daten muss das Modell für die Verwendung zusammengesetzt sein, um diese Anforderungen zu erfüllen. Diese Anforderungen bilden den Ausgangspunkt für den nächsten Schritt.

Im Schritt Verständnis der Daten (Data Understanding) werden die zuvor aufgestellten Anforderungen aus der Daten Perspektive betrachtet. Es wird geklärt welche Daten für die

Erfüllung der Anforderungen benötigt werden und ob diese Daten in notwendiger Qualtät vorliegen. Wenn die Daten nicht in notwendiger Qualität vorliegen, wird geprüft, ob die fehlenden Daten erfasst oder eingeholt werden können oder ob die Qualität der Daten an sich verbessert werden kann. Alternativ kann auch zurück zum vorhergehenden Schritt gewechselt werden, um die Anforderungen zu Überdenken und zu adaptieren (Witten, Frank, Hall, & Pal, 2016). Chapman, et al. (2000) haben in diesem Schritt auch eine erste Analyse der Daten vorgesehen. So soll hier analysiert werden ob, die erhobenen Daten zum einen überhaupt für die Erfüllung der Anforderungen relevant sind und zum anderen, ob weitere Daten benötigt werden. In diesen Analysen, sollen durch den Einsatz von Abfragen, Visualisierungstechniken und Reports ein Überblick über die erhobenen Daten erstellt werden. Damit können erste Trends erkannt und auch falsche Fährten aufgedeckt werden, welche später im Data Mining Modell zu falschen Schlüssen führen könnten.

Die festgelegten Datenquellen und die Ergebnisse der ersten Analyse fließen in den Schritt der Datenaufbereitung (Data Preparation) ein. Ziel dieses Schrittes ist es, die erhobenen Rohdaten in eine Datenbasis für die Modellerstellung zu transformieren. Dazu zählen die Aktivitäten der Auswahl der zu verwendenten Tabellen, Spalten, Attribute sowie die Transformation von Werten oder die Bereinigung der Datenbestände (Shearer, 2000). Chapman et al. (2000) beschreiben hier im Detail fünf Schritte die in folgender Reihenfolge bearbeitet werden müssen:

- Auswahl der Daten: Auswahl der Daten, welche für den Aufbau des Modells benötigt werden
- Bereinigen der Daten: Bereinigen von nicht vollständigen Datensätzen, Setzen von Standardwerten und die Interpollation von fehlenden Werten
- Aufbauen der Daten: Transformation von Daten in ein anderes Format, Durchführung von benötigten Normalisierungs- oder Denormalisierungsschritten.
- Integration der Daten: Daten aus unterschiedlichen Datenquellen werden so zusammengeführt, dass sie für die Modellierung einfach handzuhaben sind.
- Formatieren der Daten: Formatadaptierungen der Daten welche für das Modellierungswerkzeug benötigt werden. Dieser Schritt ändert nur das Format der Daten, nicht den Inhalt.

Die aufbereiteten Daten bilden die Grundlage für die Modellierung (Modeling). Dieser Schritt ist laut Witten, Frank, Hall und Pal (2016) der Schritt, in dem das eigentliche Data Mining passiert, also der Schritt, in dem das Data Mining Verfahren auf die Daten angewendet wird und das Data Mining Modell entsteht. Wie bereits zuvor erwähnt, ist es hier auch möglich auf eine Kombination aus mehreren Data Mining Verfahren zurückzugreifen. Im Detail sehen Chapman et al. (2000) in diesen Schritt auch die Auswahl der zu verwendenden Verfahren und das eigentliche modellieren des Data Mining Modells. Sie beschreiben auch, dass in diesem Schritt mehr als ein Modell erzeugt werden kann, diese Modelle können dann verglichen und schlussendlich das beste ausgewählt werden.

Das ausgewählte Modelll wird anschließend im Schritt Evaluierung (Evaluation) evaluiert. Das Kernstück dieses Schrittes ist es, die Präzission des Modells zu ermitteln, in dem das Modell mit

Testdaten evaluiert wird (Witten, Frank, Hall, & Pal, 2016). Shearer (2000) sieht in diesem Schritt die Entscheidung, ob das Modell verwendet werden kann oder ob eine weitere komplette Iteration des Prozesses angewandt werden muss, weil das Modell die Anforderungen noch nicht gut genug erfüllt.

Der letzte Schritt im Prozess ist die Verwendung (Deployment). Hier geht es darum, dass das Modell in einer Art und Weise zur Verfügung gestellt werden muss, damit es verwendet werden kann. Dieser Schritt kann auch das etablieren eines wiederholbaren Data Mining Prozesses zur kontinuierlichen Verbesserung des Modells beinhalten (Chapman, et al., 2000).

Dieser Prozess stellt die Grundlage für den in der vorliegenden Arbeit angewandten Data Mining Prozess dar und wird für die Umsetzung des Prototypen angewandt. Wenn in der weiteren Arbeit der Begriff Data Mining Prozess verwendet wird, ist dieser Prozess damit gemeint. Die ersten drei Schritte sind sehr domänenspezifisch, aus diesem Grund wird auf diese erst im Zuge der Erstellung des Prototypen im Detail eingegangen. Auf die Grundlagen der Datenstrukturen und den Umgang mit den Daten selbst wird in Abschnitt 2.3.1 näher eingegangen. Das eigentliche generieren der Data Mining Modelle unterscheidet sich in den einzelnen Data Mining Verfahren, auf diese wird in Kapitel 3 näher eingegangen. Die grundsätzlichen Möglichkeiten zur Evaluierung der Präzision eines Data Mining Modells wird in Abschnitt 2.3.2 behandelt. Abschließend wird in Abschnitt 2.3.3 darauf eingegangen wie das Data Mining Modell aktualisiert werden kann.

2.3 Data Mining: Konzepte und Techniken

Der zuvor beschriebene Data Mining Prozess stellt einen groben Überblick über die notwendigen Aktivitäten und deren Reihenfolge in Data Mining Projekten dar. Diese Informationen werden in diesem Abschnitt um tiefer gehende Themen ergänzt, welche für die Anwendung des Prozesses, im Prototyp der vorliegenden Arbeit, relevant sind. Zuerst wird auf die Strukturen und Daten für die Anwendung in Data Mining Prozessen eingegangen. Abschließend werden die, für die vorliegende Arbeit sehr wichtigen Themen: Evaluierung von erstellten Data Mining Modellen, sowie Aktualisierung von Data Mining Modellen detailliert behandelt.

2.3.1 Daten für das Data Mining

Bramer (2013) fasst die Begrifflichkeiten rund um das Thema Daten im Kontext von Data Mining zusammen. Er definiert die Namen Datensatz als Gruppierungsbegriff für alle, für das Data Mining zur Verfügung stehenden und relevanten Daten. Datensätze beinhalten viele Instanzen mit jeweils mehreren Attributen, die zur jeweiligen Instanz gehören. Eine Instanz stellt hier das Objekt dar, welches für die Data Mining Anwendung relevant ist. Die Attribute sind die Eigenschaften, mit welchen dieses Objekt beschrieben werden kann. Die Attribute der Instanzen werden im Zuge des Data Minings ausgewertet. Ein, oder mehrere, dieser Attribute

sollen für zukünftige Instanzen vorhergesagt werden. Diese Attribute werden als Klassen bezeichnet.

Laut Bramer (2013) unterscheidet man im Kontext von Data Mining zwischen zwei grundsätzlichen Arten von Attributen: kategorische Attribute und fortlaufende Attribute. Kategorische Attribute können ordinale, binäre, oder nominale Werte enthalten. Fortlaufende Attribute enthalten Zahlenwerte im Sinne von Intervallen, Messwerten oder Verhältnissen. Bei der Auswahl von einem Data Mining Verfahren müssen die in den Daten vorhandenen Arten von Attributen berücksichtigt werden. Das ausgewählte Data Mining Verfahren muss mit der entsprechenden Art von Attribut umgehen können.

Witten, Frank, Hall und Pal (2016) treffen hier die gleichen Schlüsse wie Bramer (2013), sie argumentieren aber zusätzlich, dass es sehr wichtig ist, alle Attribute einer Instanz zusammenzuziehen und relationale Strukturen zu denormalisieren. Somit sollen die Attribute, wenn sie von verschiedenen Systemen stammen, zuerst in ein System zusammengezogen und anschließend, wenn notwendig, denormalisiert werden. Ein Datensatz welcher alle Attribute der Instanzen beinhaltet und diese denormalisiert abbildet, bietet die besten Eigenschaften für das Data Mining

Fehlende Attributwerte stellen in der Vorbereitung ein Problem dar, welches gelöst werden muss. Die wenigsten Data Mining Verfahren können ohne Seiteneffekte oder Probleme mit fehlenden Werten umgehen. Wenn fehlende Attributwerte vorhanden sind, müssen die mit Ersatzwerten versorgt werden oder die fehlenden Instanzen aus dem Datensatz ausgeschlossen werden. Das Ersetzen der fehlenden Werte durch Ersatzwerte ist ein Prozess welcher gut durchdacht sein muss, es kann hier sehr leicht zu Seiteneffekten durch diese Ersetzungen kommen. (Pyle, 1999)

Attribute, welche im ganzen Datensatz keinen oder nur einen einzigen Attributwert aufweisen können und sollen aus dem Datensatz entfernt werden. Diese Attribute bieten dem Data Mining Verfahren keinen Mehrwert und neigen eher die Ergebnisse zu verfälschen. (Pyle, 1999)

In der vorliegenden Arbeit wird die Definition der Begrifflichkeiten, wie sie Bramer (2013) zusammengefasst, hat übernommen. Ein Datensatz bildet den Übergriff für eine Liste von Instanzen. Die Instanzen sind das Objekt, um welches es im Data Mining geht. Das Data Mining Modell wird mit bekannten Instanzen generiert, um Vorhersagen über zukünftige Instanzen treffen zu können. Eine Instanz wird durch seine Attribute beschrieben und im Kontext der vorliegenden Arbeit wird versucht, neue Instanzen anhand seiner Attribute, mit Hilfe des trainierten Data Mining Modells, zu klassifizieren. Data Mining Verfahren für die vorliegende Arbeit, müssen in der Lage sein, kategorische Attribute vorherzusagen. Wenn die Anzahl der Instanzen mit fehlenden Werten sehr gering ist, sollten diese Instanzen ausgeschlossen werden und auf ein ersetzen dieser fehlenden Werte durch Ersatzwerte verzichtet werden.

2.3.2 Evaluierung von Data Mining Modellen

Das generierte Data Mining Modell wird im CRISP-DM Prozess nach dem trainieren evaluiert. Han, Kamber und Pei (2012) beschreiben in diesem Schritt die Vorhersageleistung des Modells

als zu überprüfenden Faktor. Das Data Mining Modell im Prototyp der vorliegenden Arbeit ist darauf ausgelegt, eine Klassifizierung vorauszusagen. Für die Überprüfung eines Data Mining Modells, welches Klassifizierungen vorhersagt, beschreiben Han, Kamber und Pei (2012) die Konfusionsmatrix als gut geeignetes Werkzeug. In dieser Matrix wird die Präzision der Vorhersage in Form von Richtig und Falsch vorhergesagten Klassen ermittelt, in dem die Richtig und Falsch vorhergesagten Klassen den in den Daten tatsächlich vorhandenen Klassen gegenübergestellt werden. Als Kennzahl für diese Aussage werden die Kennzahlen True Positive (TP) und False Positive (FP) verwendet. Die erste Kennzahl repräsentiert die Prozentzahl der Instanzen, bei welchen eine Klasse Korrekterweise vorhergesagt wurde. Die zweite Kennzahl repräsentiert die Prozentzahl der Instanzen, bei welchen eine Klasse Fälschlicherweise vorhergesagt wurde Han, Kamber und Pei (2012).

Zur Evaluierung selbst, werden dem Modell Daten zur Vorhersage vorgelegt, bei denen das Ergebnis, außerhalb des Data Mining Modells, bereits bekannt ist. Für die Summe der vorgelegten Daten wird anschließend ausgewertet, wie viele Vorhersagen oder Klassifizierungen korrekt getroffen wurden und wie viele nicht. Ein sehr wichtiger Aspekt dabei ist, dass die Daten, die für das Training des Data Mining Modelles verwendet wurden, nicht für die Evaluierung dieses verwendet werden dürfen. Dabei muss aber trotzdem sichergestellt sein, dass sowohl die Daten, welche für das Training, als auch die Daten, welche für die Evaluierung verwendet werden, repräsentative Teilmengen derselben Problemstellung darstellen. Wenn diese beiden Bedingungen nicht erfüllt werden, kann das Ergebnis der Evaluierung sehr leicht falsche Aussagen über die Vorhersageleistung des evaluierten Data Mining Modells liefern. In der Praxis spricht man daher von zwei Datenbeständen. Einem Trainingsdatenset und einem Evaluierungsdatenset. (Bramer , 2013)

Witten, Frank, Hall und Pal (2016) beschreiben, dass beide Bedingungen leicht zu erfüllen sind, wenn die vorhandenen Datenmenge groß genug ist. In diesem Fall kann man die Datenmenge einfach, basierend auf Zufall, in zwei Teile aufteilen und einen Teil für das Training und den anderen Teil für die Evaluierung verwenden, die Methode nennt man die Hold-Out Evaluierung. In der Praxis ist die Datenmenge aber oft nicht groß genug. Aus diesem Grund wurden im Laufe der Zeit weitere Verfahren entwickelt. Sowohl Witten, Frank, Hall und Pal (2016), als auch Han, Kamber und Pei (2012) nennen die Kreuzvalidierung, die Bootstrap Methode und die Leave-One-Out Methode als wichtigste Verfahren.

Bei der Kreuzvalidierung wird die gesamte Datenmenge in N etwa gleich große Partitionen zerlegt. Das Training und die Evaluierung des Modells wird n mal durchgeführt. Dabei werden immer N-1 Partitionen für das Training und eine Partition für die Evaluierung verwendet. Der Durchschnittswert des Ergebnisses über alle Wiederholungen wird anschließend verwendet (Witten, Frank, Hall, & Pal, 2016). Han, Kamber und Pei (2012) beschreiben, dass in der Praxis oft 10 Partitionen zum Einsatz kommen. Sie beschreiben auch, dass die Leave-One-Out Methode ein Sonderfall der Kreuzvalidierung ist, bei dem die Datenmenge in so viele Partitionen aufgeteilt wird, wie Instanzen im Datensatz vorhanden sind. So wird eine jede Instanz einmal als Testdatensatz verwendet, während in der jeweiligen Iteration alle anderen Einträge zum Trainieren verwendet wurden. Der Nachteil dieses Verfahrens ist der hohe Aufwand für die große Anzahl an Wiederholungen. Bei der Bootstrap Methode zur Evaluierung,

welche von Efron (1979) beschrieben wurde, wird das statistische Prinzip des Bootstrappings verwendet, um Trainings- und Evaluierungsdatensätze zu bilden. Dazu werden mehrere Gruppen aus Trainingsdatensätzen und Evaluierungsdatensätzen durch das Ziehen von Instanzen aus der Gesamtmenge gebildet. Bei dieser Methode werden die Instanzen nach dem ziehen wieder zurückgelegt, wodurch die Möglichkeit entsteht, dass einzelne Instanzen auch mehrfach innerhalb einer Partition vorkommen können. Witten, Frank, Hall und Pal (2016) merken an, dass diese Methode schlechter ist als eine Kreuzvalidierung, da bei der am häufigsten eingesetzten Variante dieses Verfahren, der 0.632 Bootstrap Methode, nur ungefähr 63% der Daten für die Evaluierung verwendet werden können. Bei der Kreuzvalidierung mit zehn Iterationen, werden hingegen 90% der Daten, auch für die Evaluierung verwendet.

Die Evaluierung in der vorliegenden Arbeit wird mit Aufstellung einer Konfusionsmatrix und der Berechnung der Präzision des Modells durchgeführt. Für die Aufteilung der Daten, in Trainings- und Evaluierungsdaten, wird eines der genannten Verfahren verwendet. Welches, wird im Zuge der Umsetzung des Prototyps auf Basis der Anzahl der Daten und der Laufzeit für eine Trainingsiteration entschieden. Diese Evaluierung wird auch zur laufenden Prüfung des verwendeten Modells zur Erhebung der Werte, welche für die Beantwortung der Forschungsfrage benötigt werden, verwendet.

2.3.3 Aktualisieren von Data Mining Modellen

Zur kontinuierlichen Verbesserung von Data Mining Modellen während es in Betrieb ist, wurden in der Literatur zwei grundsätzliche Möglichkeiten gefunden. Zum einen kann das Modellieren, Evaluieren und Ausliefern im klassischen Data Mining Prozess nach Chapman (2000) automatisiert und wiederholt mit neuen Datenständen durchgeführt werden. Der Prozess wurde grundsätzlich als iterativer Prozess vorgesehen. Wenn sich jedoch die grundsätzliche Struktur der Daten ändert ist eine einfache neuerliche Generierung des optimierten Modells mit neuen Daten nicht mehr einfach möglich. Das neue Data Mining Modell wird nach dem neuen Training automatisch durch das aktuelle ersetzt. Zum anderen gibt es aber auch spezialisierte Data Mining Verfahren, welche ein kontinuierliches Trainieren des Data Mining Modells durch Verarbeitung eines laufenden Datenstroms erlauben. Dieses Data Mining Verfahren sind aber auf sehr große Datenmengen spezialisiert, bei denen ein Trainieren des Data Mining Modells mit der gesamten Datenbasis aufgrund der großen Datenmenge nicht mehr sinnvoll möglich ist (Gaber, Zaslavsky, & Krishnaswamy, 2010).

Gaber, Zaslavsky und Krishnaswamy (2010) stellen Kriterien auf, welche es ermöglichen zu identifizieren, ob für einen spezifischen Anwendungsfall ein Data Mining Verfahren, welches auf einem Datenstrom operieren kann sinnvoll ist, oder, ob die klassische Variante eingesetzt werden soll. Die für die vorliegende Arbeit relevanten Kriterien sind:

- Anforderung an eine Echtzeitanalyse von neuen Daten
- Geschwindigkeit und Größe der neu generierten Datenmengen (In Relation zur Verfügung stehenden Rechenleistung)
- Speicherbedarf der entstehenden Daten

Für die vorliegende Arbeit gibt es keine Anforderung an eine Echtzeitanalyse von neuen Daten. Es ist ausreichend wenn neue Daten erst zyklisch mit Zeitverzögerung in das Modell integriert werden. Eine Analyse des für die Evaluierung in der vorliegenden Arbeit eingesetzten Smart Homes hat ergeben, dass pro Jahr circa ein Gigabyte an Rohdaten erfasst werden. Diese Datenmenge umfasst aber sämtliche durch Sensoren erfasste Daten und somit auch Daten, die nicht für die Beleuchtungssteuerung relevant sind. Dabei gibt es pro Tag im Durchschnitt 12.000 Ereignisse die protokolliert werden. Gaber, Zaslavsky und Krishnaswamy (2010) nannten als Beispiele für datenstromverarbeitende Data Mining Systeme, Systeme bei denen täglich mit mehreren hundert Gigabyte an neuen Daten umgegangen werden muss. Da die drei Kriterien keine klare Argumentation in Richtung datenstrombasiertes Data Mining zulassen und da es in ähnlichen Arbeiten (Das, Cook, Battacharya, Herman, & Lin, 2002), bei denen Data Mining Verfahren im Kontext eines Smart Homes eingesetzt wurden, ebenso nicht notwendig war auf datenstrombasiertes Data Mining zurückzugreifen, wird auch in der vorliegenden Arbeit keine Notwendigkeit dafür gesehen.

Es wurden zwei grundsätzliche Möglichkeiten identifiziert eine laufende Verbesserung eines sich in Betrieb befindlichen Data Mining Modells zu realisieren. Ein zyklischer Neuaufbau des gesamten Modells mit den alten und den neuen Trainingsdaten, sowie die Möglichkeit ein Data Mining Modell laufend, basierend auf einem kontinuierlichen Datenstrom, weiter zu trainieren. Die zweite Variante ist für sehr große Datenmengen gedacht und ist mit höherem Aufwand verbunden, als die erste Variante. Es wurden für die vorliegende Arbeit keine Anforderungen erkannt, welche eine Umsetzung des Prototyps mit einem datenstrombasierten Data Mining Verfahren erfordern würde. Aus diesem Grund wird in der vorliegenden Arbeit die erste Variante eingesetzt und das Data Mining Modell zyklisch komplett neu aufgebaut und in die Anwendung integriert.

2.4 Zusammenfassung

In diesem Kapitel wurde ein Überblick über die Anwendung von Data Mining geschaffen und die Begriffe Data Mining, Machine Learning und Smart Home für die vorliegende Arbeit definiert. Es wurde auf die notwendigen Schritte für die Durchführung eines Data Mining Projektes, sowie auf einen standardisierten Data Mining Prozess eingegangen und aus den gewonnenen Erkenntnissen Einschränkungen für die vorliegende Arbeit abgeleitet.

Als Smart Home wird in der vorliegenden Arbeit eine Wohneinheit verstanden, welche mit vernetzten Sensoren und Aktoren ausgestattet ist. Dieses Smart Home wird um eine künstliche Intelligenz zur Automatisierung der Beleuchtungssteuerung, auf Basis eines Data Mining Modells, erweitert.

Der Begriff Data Mining wurde für die vorliegende Arbeit als Prozess zur Wissensgenerierung aus einer großen Datenmenge durch Analyse dieser, auf das Vorhandensein von interessanten Mustern definiert. Dabei wird unter Data Mining der gesamte Prozess verstanden, welcher notwendig ist, um aus einer großen Datenmenge nützliches Wissen zu generieren. Als Data Mining Verfahren wird in der vorliegenden Arbeit ein konkretes Verfahren zur Generierung von

Wissen aus diesen Daten definiert. Die für die Wissensgenerierung relevanten Data Mining Verfahren kommen aus den Bereichen der Statistik, der künstlichen Intelligenz und dem Machine Learning. Da sich in den Ansätzen aus dem Bereich des Machine Learnings viele Ideen aus der Statistik und der künstlichen Intelligenz wiederfinden, wurde der Fokus für die vorliegende Arbeit auf Machine Learning gelegt. Im Zuge der Definition von Data Mining, wurde für die vorliegende Arbeit eine Einschränkung auf Data Mining Verfahren zur Lösung von Klassifizierungsproblemen definiert, da es sich im Anwendungsfall der vorliegenden Arbeit, genau um ein solches Problem handelt.

Der Begriff Machine Learning, wurde in der vorliegenden Arbeit als die Anwendung einer konkreten Technik zur Extraktion von Wissen aus Daten definiert. Darunter wird der konkrete Algorithmus verstanden, welcher es ermöglicht, Wissen aus Daten zu generieren. Machine Learning selbst ist somit ein wichtiger Prozessschritt im übergeordneten Data Mining Prozess. In der vorliegenden Arbeit wurden Machine Learning Algorithmen als eine Art von Data Mining Verfahren definiert. In weiterer Folge wird der Überbegriff Data Mining Verfahren für solche Algorithmen und Techniken verwendet. Es wurde auch eine Einschränkung auf Techniken, aus dem Bereich des beaufsichtigten Lernens definiert.

Der Data Mining Prozess nach Chapman et al. (2000), welcher in diesem Kapitel beschrieben wurde, stellt ein standardisiertes Vorgehen zur Anwendung von Data Mining dar. Dieser Prozess beschreibt ein iteratives Vorgehen, welches die Schritte: fachliches Verständnis, Verständnis der Daten, Datenaufbereitung, Modellierung, Evaluierung und Verwendung beinhaltet. Das eigentliche Data Mining Verfahren, kommt dabei im Schritt Modellierung zur Anwendung. Dieser Prozess wurde als Vorlage für die Umsetzung des Prototyps in der vorliegenden Arbeit definiert.

Für die Aktualisierung von Data Mining Modellen wurden verschiedene Varianten aus der Literatur recherchiert und analysiert. Für den vorliegenden Anwendungsfall wurde definiert, dass ein zyklischer Neuaufbau des Data Mining Modells ausreichend ist. Der Einsatz eines, wesentlich aufwendigeren, datenstrombasierten, Data Mining Verfahrens oder die Einschränkung auf Data Mining Verfahren, welche auch eine laufende Aktualisierung eines trainierten Data Mining Modells zulassen, wurde als nicht notwendig eingestuft. Diese Entscheidung wurde damit begründet, dass die Datenmenge und die Laufzeit im vorliegenden Anwendungsfall klein genug sind, um einen vollständigen Neuaufbau der Modelle in akzeptablen Zeitraum durchzuführen.

Zusammengefasst wurden in diesem Kapitel die grundlegenden Begriffe rund um das Thema Data Mining zur vorliegenden Arbeit definiert, ein Überblick über die Anwendung von Data Mining geschaffen und alle Informationen aufbereitet, um es im folgenden Kapitel zu ermöglichen, auf konkrete Data Mining Verfahren für die vorliegenden Arbeit einzugehen und eines daraus auszuwählen.

3 DATA MINING VERFAHREN

In diesem Kapitel werden verschiedene Data Mining Verfahren, entsprechend der im Vorkapitel getroffenen Einschränkung auf Klassifizierungsverfahren, erläutert und analysiert. Ziel dieser Art von Data Mining Verfahren ist es, neue Datensituationen auf Basis des trainierten Data Mining Modells zu klassifizieren, um Vorhersagen über die neue Datensituation treffen zu können. Zu Beginn des Kapitels werden grundsätzliche Data Mining Verfahren für Klassifizierungsaufgaben, welche mittels einer Literaturrecherche in Fachbüchern, zum Thema Data Mining, gefunden wurden, aufgelistet und ausgewählt. Anschließend werden die gefundenen Data Mining Verfahren auf für die vorliegende Arbeit relevante Data Mining Verfahren eingeschränkt und auf diese Auswahl näher eingegangen. Konkret wird für jedes untersuchte Data Mining Verfahren auf die grundsätzliche Funktionsweise eingegangen und eine Klassifizierung der wichtigsten Eigenschaften durchgeführt. Im Anschluss werden Anforderungen an ein Klassifizierungsverfahren für den Anwendungsfall der vorliegenden Arbeit erörtert, um die Basis für die Auswahl eines Data Mining Verfahrens zu bilden. Der untersuchte Anwendungsfall, sowie die Anforderungen an diesen, bilden im Abschluss die Grundlage für die Auswahl eines Data Mining Verfahren für die Umsetzung des Prototyps in der vorliegenden Arbeit.

3.1 Arten von Data Mining Verfahren für Klassifizierungsaufgaben

In diesem Abschnitt werden die, in der Literatur gefundenen, Data Mining Verfahren für Klassifizierungsprobleme aufgelistet und für die detaillierte Betrachtung ausgewählt. In der Literatur wurde speziell nach Arbeiten gesucht, welche sich allgemein mit dem Thema der klassifizierenden Data Mining Verfahren beschäftigen und konkrete Verfahren dazu beschreiben oder auflisten. Für die Recherche wurde primär auf Fachbücher zurückgegriffen, welche einen Überblick über bestehende Data Mining Verfahren geben und deren Vor- und Nachteile behandeln. Ergänzt wurde diese Recherche durch eine Recherche nach Vergleichen und Auflistungen von weit verbreiteten Data Mining Verfahren auf Google Scholar.

Wie in Tabelle 1 dargestellt, wurden die Arbeiten von Shearer (2000), Wu et. al (2007), Han, Kamber und Pei (2012), Bramer (2013) und Kotu und Despande (2015) im Zuge der Literaturrecherche untersucht. Darin wurden die Data Mining Verfahren „lineare Modelle“, Regelinduktion, Entscheidungsbäume, „neuronale Netwerke“, „nächste Nachbarn Klassifikation“, „fallbasiertes Schließen“, „genetische Algorithmen“, „Naive Bayes Klassifikation“ sowie Stützvektormaschinen als für Klassifizierungsaufgaben eingesetzte Verfahren identifiziert.

In der Arbeit von Wu et. al (2007) wurden die zehn in Publikationen meist verwendeten Data Mining Algorithmen identifiziert. Darunter befinden sich fünf Algorithmen für

Klassifizierungsaufgaben: der C4.5 Algorithmus, die nächste Nachbarn Klassifikation, die naive Bayes Klassifizierung, die Stützvektormaschinen und CART. Die konkreten Algorithmen C4.5 und CART sind Algorithmen welche auf dem Data Mining Verfahren der Entscheidungsbäume aufbauen. Zusammengefasst lässt sich damit sagen, dass diese vier Data Mining Verfahren eine weite Verbreitung und Anwendung finden.

Data Mining Verfahren	Aufgelistet von
Lineare Modelle	Shearer (2000)
Regelinduktion	Shearer (2000), Kotu und Despande (2015), Bramer (2013), Han, Kamber und Pei (2012)
Entscheidungsbäume	Shearer (2000), Kotu und Despande (2015) Bramer (2013), Han, Kamber und Pei (2012), Wu et. al (2007)
Neuronale Netzwerke	Shearer (2000), Kotu und Despande (2015), Han, Kamber und Pei (2012)
Nächste Nachbarn Klassifikation	Shearer (2000), Kotu und Despande (2015), Bramer (2013), Han, Kamber und Pei (2012), Wu et. al (2007)
Fallbasiertes Schließen	Shearer (2000), Han, Kamber und Pei (2012)
Genetische Algorithmen	Shearer (2000), Kamber und Pei (2012)
Naive Bayes Klassifikation	Kotu und Despande (2015), Bramer (2013), Han, Kamber und Pei (2012), Wu et. al (2007)
Stützvektormaschinen	Kotu und Despande (2015), Han, Kamber und Pei (2012), Wu et. al (2007)

Tabelle 1: Auflistung von Klassifizierungsverfahren aus der Literaturrecherche

Han, Kamber und Pei (2012) beschreiben bei den „genetischen Algorithmen“, dass diese im Kontext von Data Mining zwar für Klassifizierungsaufgaben eingesetzt werden können, dass sie in der Praxis aber primär zur Evaluierung von anderen Algorithmen eingesetzt werden und nicht für die Klassifizierungsaufgabe selbst. Genetische Algorithmen haben ihre Stärke in sehr speziellen Anwendungsgebieten und finden keine weite Verbreitung in Data Mining Projekten. Aus diesen Gründen werden genetische Algorithmen in der vorliegenden Arbeit nicht weiter behandelt. Han, Kamber und Pei (2012) sehen beim Data Mining Verfahren „fallbasiertes Schließen“ eine große Problematik in der Automatisierung des Trainings des Data Mining Modells. Da diese Automatisierung laut Han, Kamber und Pei (2012) noch ein sehr offenes Gebiet in der Forschung ist wird dieses Data Mining Verfahren in der vorliegenden Arbeit nicht weiter behandelt.

Zusammengefasst wurden die folgenden Data Mining Verfahren, für Klassifizierungsaufgaben in der Literatur gefunden, welche für die vorliegende Arbeit weiter betrachtet werden:

- Entscheidungsbäume (Englisch: decision trees)
- Stützvektormaschinen (Englisch: support vector machine)
- Nächste Nachbarn Klassifikation (Englisch: k-nearest neighbor classification)
- Naive Bayes Klassifikation (Englisch: naïve bayes classification)

- Neuronale Netzwerke (Englisch: neural network)
- Regelinduktion (Englisch: rule induction)
- Lineare Modelle zur Klassifikation (Englisch: linear models for classification)

Die Reihenfolge wurde so gewählt, dass zuerst die Verfahren, welche laut Wu et. al. (2007) zu den zehn wichtigsten zählen, untersucht werden und anschließend die restlichen Verfahren. Da der Großteil der Literatur zu diesen Themen auf Englisch vorliegt, wurden in der Auflistung auch die Englischen Begriffe zur einfacheren Übersicht hinzugefügt. In den folgenden Abschnitten wird auf die aufgelisteten Data Mining Verfahren im Detail eingegangen.

Kotu und Deshpande (2015) verglichen Data Mining Verfahren miteinander und definierten relevante Kriterien, die für den Vergleich von Data Mining Verfahren zur Auswahl für einen konkreten Anwendungsfall relevant sind.

- Struktur des Data Mining Modells: Wie sieht das Modell aus?
- Eingabeattribute: Wie sind die Eingabeattribute für das Modell definiert?
- Ausgabeattribute: Wie sind die Ausgabeattribute für das Modell definiert?
- Vorteile: Welche allgemeinen Vorteile hat das Data Mining Verfahren?
- Nachteile: Welche allgemeinen Nachteile hat das Data Mining Verfahren?

Diese Kriterien werden als Basis für die Auswahl von Data Mining Verfahren in der vorliegenden Arbeit verwendet und für alle untersuchten Data Mining Verfahren in den folgenden Abschnitten abgeleitet.

3.2 Entscheidungsbäume

Das Data Mining Verfahren der Entscheidungsbäume wurde von Quinlan (1986) vorgestellt. Er beschreibt damit ein Verfahren bei welchem, während dem Training des Data Mining Modells, hierarchische Strukturen in Form von Bäumen abgeleitet werden. Jeder Knoten dieses Baumes entspricht einer Entscheidung welche auf Basis der Attribute einer Instanz im Datensatz getroffen wird. Bei der Vorhersage der Klassifizierung für eine neue Instanz wird der Entscheidungsbaum, ausgehend von seinem Wurzelknoten, durchlaufen und die Entscheidung eines jeden Knoten ausgewertet. Eine jede Entscheidung verfolgt dabei das Ziel, einer Vorhersage zu einer Klassifizierung der Instanz einen Schritt näher zu kommen. Ist man in einem Ast des Baumes angelangt an dem es keine weitere Entscheidung gibt wird in diesem Knoten die entsprechende Vorhersage der Klassifizierung gefunden. Man spricht hier von einer teile und herrsche Strategie weil mit jeder Entscheidung die Menge der Möglichkeiten kleiner wird. Der Algorithmus von Quinlan (1986) arbeitet hier mit der Entropie der Attribute im Trainingsdatensatz. Mit der Entropie kann errechnet werden welches Attribut den Datensatz mit dem größten Informationsgewinn aufteilt. Neuere Implementierungen von Endscheidungsbaum Algorithmen verwenden hier auch andere statistische Kennzahlen, wie den Gini Koeffizienten (Han, Kamber, & Pei, 2012). Für die Berechnung des Attributes wird der gesamte

Trainingsdatensatz verwendet. Bei diesem Data Mining Verfahren gibt es unterschiedliche Ansätze, wie bei der Auswahl der Attribute für Entscheidungen vorgegangen wird und wie der Baum strukturiert wird. Diese zwei Punkte stellen auch die Herausforderung bei der Implementierung dieses Data Mining Verfahrens dar (Witten, Frank, Hall, & Pal, 2016). Der von Quinlan (1986) vorgestellte Algorithmus ID3 war nur in der Lage eine von zwei Klassen vorherzusagen und konnte somit nur eine boolesche Klassifizierung vornehmen. Im Laufe der Zeit wurden aber neue, bessere Algorithmen entwickelt, welche auch mit mehreren Klassen umgehen können (Witten, Frank, Hall, & Pal, 2016). Einer der wichtigsten Vertreter davon ist der bereits zuvor erwähnte C4.5 Algorithmus, welcher von Quinlan (1993) vorgestellt wurde.

In Folge wird dieses Data Mining Verfahren anhand der Kriterien nach Kotu und Deshpande (2015) klassifiziert:

Modell: Das Modell besteht bei Entscheidungsbäumen aus einem Baum, wessen Knoten Entscheidungsregeln für die Vorhersage von Klassifizierungen bilden und den Baum dabei in immer kleinere Partitionen teilen (Quinlan, 1986).

Eingabeattribute: Entscheidungsbäume können mit unterschiedlichen Typen von Attributen umgehen. Es gibt keine Restriktion auf bestimmte Typen von Attributen (Kotu & Deshpande, 2015).

Ausgabeattribute: Die vorherzusagende Klasse kann kein fortlaufender numerischer Wert sein. Es sind nur Klassifizierungen möglich (Kotu & Deshpande, 2015).

Vorteile: Quinlan (1986) beschreibt, dass Entscheidungsbäume den großen Vorteil haben, dass sie sehr gut visualisiert werden können und auch Laien anhand dieser Visualisierung die Funktionsweise einfach nachvollziehen können. Zusätzlich sehen Kotu und Deshpande (2015) den Vorteil, dass keine Umwandlung der Attribute und der vorhergesagten Klasse notwendig ist da dieses Data Mining Verfahren mit klassifizierenden und numerischen Attributen umgehen kann.

Nachteile: Kotu und Deshpande (2015) fassen zusammen, dass Entscheidungsbäume leicht dazu neigen, sich zu sehr an die Trainingsdaten anzupassen, dass kleine Änderungen in den Trainingsdaten zu komplett anderen Entscheidungsbäumen führen können und dass es sehr herausfordernd sein kann, die richtigen Entscheidungsknoten abzuleiten.

Im Laufe der letzten Jahrzehnte haben sich unterschiedliche Techniken und Methoden entwickelt mit welchen die Erstellung von Entscheidungsbäumen optimiert und beeinflusst werden kann (Witten, Frank, Hall, & Pal, 2016). Das größte Problem von Entscheidungsbäumen ist Überanpassung des erstellten Baumes an den Trainingsdatensatz. Dadurch sind die Vorhersagen für den Trainingsdatensatz sehr gut, aber für neue Daten sehr schlecht. Bei einer Überanpassung ist das Data Mining Modell so stark auf den Trainingsdatensatz angepasst, dass es nur mehr für die dort enthaltenen Fälle gute Vorhersagen treffen kann. Ein Entscheidungsbaum hat grundsätzlich keine Grenzen in Tiefe oder Breite des Baumes, durch einen zu tiefen oder zu breiten Baum entsteht sehr leicht einer Überanpassung. Aus diesem

Grund wurde die Technik des Pruning (Ausästen[5]) entwickelt. Beim Pre-Pruning wird der Baum in Tiefe und Breite nach vorgegebenen Regeln eingeschränkt. Hier passiert das Pruning während der Entscheidungsbaum aufgebaut wird, es wird entschieden ob für eine Entscheidung ein weiterer Knoten eingeführt werden darf oder nicht. Beim Post-Pruning wird der originale Entscheidungsbaum nach der Erstellung beschnitten und Teilbäume auf Basis von Regeln entfernt, um die Größe des Entscheidungsbaumes zu reduzieren (Kotu & Deshpande, 2015). Beide Varianten werden eingesetzt, um der Überanpassung des Entscheidungsbaumes vorzubeugen und so eine höhere Präzision der Vorhersagen zu erhalten.

3.3 Stützvektormaschinen

Cortes und Vapnik (1995) haben ein Data Mining Verfahren für Klassifizierungsprobleme vorgestellt, welches einen nicht linearen Problemraum zuerst auf einen linearen Lösungsraum abbildet und anschließend eine Klassifizierung durch das Ziehen von Grenzen in diesem Lösungsraum durchführt. Der Datensatz wird auf einen mehrdimensionalen linearen Raum abgebildet. Jedes Attribut mit seinen Werten bildet dabei eine Dimension im linearen Lösungsraum. Jede Instanz im Datensatz wird somit als mehrdimensionaler Vektor im Lösungsraum gesehen. Anschließend werden alle Punkte, die denselben Klassenwert eines Attributes haben, in der jeweiligen Dimension miteinander verbunden. Durch diese Verbindung ergibt sich je Klasse und Attribut eine Hülle, welche alle Punkte mit dem jeweiligen Klassenwert umhüllt. Anschließend werden Grenzen zwischen den einzelnen Hüllen berechnet, welche jeweils den maximalen Abstand zwischen zwei Klassenhüllen eines Attributes bilden. Mit diesen Grenzen können Vorhersagen getroffen werden. Der vorherzusagende Datensatz wird in den linearen Lösungsraum transformiert und auf Basis der Platzierung mit den bekannten Attributwerten im Lösungsraum und den Grenzen des vorherzusagenden Attributes, kann eine Klasse vorhergesagt werden (Cortes & Vapnik, 1995). Laut Kotu und Deshpande (2015) ist dieses Data Mining Verfahren für Mustererkennung und Text Mining sehr gut geeignet, es findet aber auch in anderen Anwendungsbereichen eine große Verbreitung, weil es in vielen Anwendungsgebieten gute Ergebnisse liefert.

In Folge wird dieses Data Mining Verfahren anhand der Kriterien nach Kotu und Deshpande (2015) klassifiziert:

Modell: Das Modell zu diesem Data Mining Verfahren wird durch Speicherung der Instanzen als Vektoren und der Grenzen zwischen den Hüllen umgesetzt. Es werden nicht alle Instanzen der Trainingsdatensätze für das Modell benötigt, nur die Vektoren, die an die Hüllen grenzen und somit die Hüllen der jeweiligen Klasse bilden, müssen im Modell vorhanden sein (Kotu & Deshpande, 2015).

[5] Auf den Begriff Ausästen wird in Folge verzichtet. Der Begriff Pruning hat sich in diesem Kontext etabliert.

Eingabeattribute: Alle Eingabeattribute müssen numerisch sein, klassifizierende Attribute müssen im linearen Lösungsraum auf fortlaufende Zahlenwerte abgebildet werden (Kotu & Deshpande, 2015).

Ausgabeattribute: Stützvektormaschinen sind für Klassifizierungsaufgaben gedacht und in diesem Bereich auch weit verbreitet. In den letzten Jahren wurden aber auch Algorithmen speziell für fortlaufende Vorhersagen mit Hilfe von Stützvektormaschinen entwickelt (Witten, Frank, Hall, & Pal, 2016).

Vorteile: Laut Han, Kamber und Pei (2012) ist das Modell dadurch, dass nur die Vektoren gespeichert werden, welche für die Hülle einer Klasse notwendig sind, sehr kompakt. Außerdem beschreiben sie, dass dieses Data Mining Verfahren sehr robust gegen Überanpassung des Modells und Ausreißer in den Daten ist. Kotu und Deshpande (2015) ergänzen, dass Stückvektormaschinen sehr gut mit nicht linearer Korrelation zwischen Attributen umgehen können.

Nachteile: Witten, Frank, Hall und Pal (2016) beschreiben die Nachteile, dass das Model von Stützvektormaschinen mit mehr als drei Dimensionen nicht vernünftig visualisierbar ist. Sie beschreiben auch, dass das Mapping des Problemraums auf einen gut geeigneten linearen Lösungsraum, oft eine aufwendige Aufgabe mit vielen Versuchen darstellt. Diesen Nachteil nennen auch Kotu und Deshpande (2015) in Kombination mit der hohen Rechenleistung, die für das Trainieren eines Modells benötigt wird.

3.4 Nächster Nachbar Klassifikation

Laut Witten, Frank, Hall und Pal (2016) handelt es sich bei diesem Data Mining Verfahren um ein Verfahren welches in den 1960er Jahren als Lösung für Klassifizierungsprobleme entdeckt und seitdem eingesetzt, sowie weiterentwickelt wurde. Bei dieser Art von Data Mining Verfahren werden im Zuge der Generierung des Modells die Trainingsinstanzen auf einen n Dimensionalen Raum abgebildet. Bei der Vorhersage von zukünftigen Instanzen wird die Entfernung der Testinstanz zu den bekannten Trainingsinstanzen in diesem n dimensionalen Raum ermittelt und so der nächste Nachbar zur Testinstanz gefunden. Für die Ermittlung der Distanzen wird in den meisten Algorithmen die Euklidische Distanzfunktion verwendet (Witten, Frank, Hall, & Pal, 2016). Johns (1961) hat dieses Vorgehen erstmalig für die Klassifizierungsprobleme evaluiert und beschrieben.

In Folge wird dieses Data Mining Verfahren anhand der Kriterien nach Kotu und Deshpande (2015) klassifiziert:

Modell: In aktuellen Implementierungen dieses Data Mining Verfahrens wird das Modell als kD-Baum gespeichert. Diese Datenstruktur, welche von Friedman, Bentley und Finkel (1977) vorgestellt wurde, ist speziell darauf ausgelegt Punkte in einem k dimensionalen Raum zu speichern und effizient wieder zu finden. Laut Witten, Frank, Hall und Pal (2016) ist es erst durch diese Datenstruktur möglich geworden, große Data Mining Modelle mit diesem Data Mining Verfahren in akzeptabler Laufzeit aufzubauen und Vorhersagen zu treffen. In dieser

Struktur wird eine jede Trainingsinstanz als Punkt mit k Dimensionen im Baum abgespeichert, Knoten im Baum trennen jeweils den Wertebereich in kleinere Partitionen auf. Bei der Vorhersage wird zuerst im Baum nach unten gegangen um den Bereich in dem die gesuchte Instanz liegt auf einen Sub-Baum einzugrenzen. Am Ende wird der nächste Nachbar in dieser Partition ermittelt.

Eingabeattribute: Laut Bramer (2013)wird dieses Data Mining Verfahren meist eingesetzt, wenn alle Attribute im Datensatz fortlaufende Werte beinhalten. Es ist aber auch möglich dieses Data Mining Verfahren mit kategorischen Werten anzuwenden, indem diese auf Zahlenwerte abgebildet werden. Auf Basis dieser Werte kann die Distanz errechnet werden. Diese Vorgehensweise bringt Probleme mit sich, weil zum Beispiel Reihenfolgen der Klassen gebildet werden, die es in der Realität nicht gibt und welche bei diesem Data Mining Verfahren das Modell verfälschen.

Ausgabeattribute: Es wird der Wert eines Attributes des nächsten gefundenen Nachbarn vorhergesagt, es können damit nur Klassen und keine fortlaufenden Werte vorhergesagt werden (Kotu & Deshpande, 2015).

Vorteile: Kotu und Deshpande (2015) fassen zusammen, dass das Data Mining Modell bei diesem Data Mining Verfahren einfach zu erstellen ist. Dieses Data Mining Verfahren kann auch gut mit fehlenden und unbekannten Werten in der Vorhersageinstanz umgehen. Witten, Frank, Hall und Pal (2016) nennen zusätzlich den Vorteil, dass das Modell jederzeit durch neue Trainingsdaten erweitert werden kann, ohne das gesamte Modell neu aufbauen zu müssen.

Nachteile: Witten, Frank, Hall und Pal (2016) beschreiben, dass dieses Data Mining Verfahren nur gut funktioniert, so lange die Anzahl der Attribute klein bleibt. Mit diesem Data Mining Verfahren hat jedes Attribut dasselbe Gewicht auf die Vorhersage. Dieses Data Mining Verfahren ist auch sehr anfällig für Ausreißer im Wertebereich. Eine kleine Anzahl von Ausreißern kann die Vorhersagequalität bereits stark beeinflussen. Kotu und Deshpande (2015) nennen die Laufzeit zur Vorhersage bei sehr großen Data Mining Modellen, sowie etwaige Probleme, welche durch die Normalisierung von klassifizierenden Attributwerten auf fortlaufende Werte entstehen können, als Nachteile dieses Data Mining Modells.

3.5 Naive Bayes Klassifikation

Die Naive Bayes Klassifikation geht auf ein von Bayes (1763) beschriebenes Verfahren zur Bestimmung der bedingten Wahrscheinlichkeit zurück. Die sogenannte einfache Bayes Klassifikation, welche in den meisten Arbeiten Naive Bayes Klassifikation genannt wird, wurde aus dem Satz von Bayes abgeleitet und stellt ein sehr einfaches und in der Praxis sehr erfolgreiches Klassifizierungsverfahren dar (Han, Kamber, & Pei, 2012).

Das Data Mining Verfahren Naive Bayes Klassifikation geht davon aus, dass jedes Attribut nur von der vorherzusagenden Klasse abhängt. Auf dieser Annahme wird für alle Attribute im Datensatz die Wahrscheinlichkeit errechnet, mit welcher die jeweiligen Klassen der Attribute bei einer bestimmten vorherzusagenden Klasse vorkommen. Mit dem Produkt dieser einzelnen

Wahrscheinlichkeiten je Attribut und zugeordneter Klasse, kann anschließend die Wahrscheinlichkeit für eine Klasse des zu vorhersagenden Attributes ermittelt werden, wenn die anderen Attribute entsprechend mit bestimmte Werten besetzt sind (Bramer , 2013).

Domingos und Pazani (1997) haben die Leistung der Naive Bayes Klassifizierung in einer empirischen Studie untersucht und mit anderen Klassifizierungsverfahren verglichen. Die Naive Bayes Klassifizierung kann vergleichbare Ergebnisse wie wesentlich aufwändigere Verfahren liefern und hat bei einer kleineren Anzahl von Instanzen in den Trainingsdatensätzen bessere Ergebnisse als andere Verfahren geliefert. Ebenso haben Domingos und Pazani (1997) in ihrer Arbeit argumentiert, dass die Naive Bayes Klassifikation auch bei Abhängigkeiten zwischen den Attributen sehr gute Ergebnisse liefern kann, wenn die Attribute nicht zu stark miteinander korrelieren.

In Folge wird dieses Data Mining Verfahren anhand der Kriterien nach Kotu und Deshpande (2015) klassifiziert:

Modell: Das Data Mining Modell zu einer Naive Bayes Klassifizierung besteht aus einer Tabelle, in welcher Wahrscheinlichkeiten und bedingte Wahrscheinlichkeiten für jedes Attribut im Zusammenhang zu einer Vorhersageklasse hinterlegt sind (Kotu & Deshpande, 2015). Das Modell entspricht damit einer relativ einfachen Abbildung in einer Software, da nur einfache Abfragen auf die Tabelle für Vorhersagen ausgeführt werden müssen, um Vorhersagen durchzuführen.

Eingabeattribute: Laut Bramer (2013) kann die Naive Bayes Klassifikation nur mit kategorischen Werten umgehen, da je Attributwert die bedingte Wahrscheinlichkeit für eine Kategorie errechnet wird. Dieses Problem kann aber umgegangen werden, in dem man fortlaufende numerische Werte in Intervalle einteilt und jedem dieser Intervalle eine Kategorie zuweist.

Ausgabeattribute: Die Naive Bayes Klassifikation ist ein typisches Klassifizierungsverfahren, welches nur für die Vorhersage von kategorischen Attributen verwendet werden kann. Es ist daher nicht möglich fortlaufende Werte mit diesem Data Mining Verfahren vorherzusagen (Bramer , 2013).

Vorteile: Domingos und Pazani (1997) fassen zusammen, dass das Naive Bayes Data Mining Verfahren ein einfaches Verfahren ist, welches sehr gute Ergebnisse liefern kann. Das Verfahren basiert sehr stark auf Statistik und Wahrscheinlichkeiten, was es einfach gestaltet, die Vorhersagen im System nachvollziehen zu können. Das Training und die Integration eines Naive Bayes Data Mining Modells, ist sehr einfach und in kurzer Zeit möglich (Bramer , 2013). Witten, Frank, Hall und Pal (2016) ergänzen hier, dass fehlende Werte bei Attributen für die Naive Bayes Klassifikation keine Probleme verursachen.

Nachteile: Bramer (2013) nennt als großen Nachteil der Naive Bayes Klassifikation, dass dieses Data Mining Verfahren schlechte Ergebnisse liefert, wenn die Verteilung der Klassen sehr ungleichmäßig ist. Wenn eine Klasse nur in sehr wenigen Instanzen vorkommt, sind die Vorhersagen für diese Klasse in vielen Fällen nicht sehr gut. Witten, Frank, Hall und Pal (2016), sowie Kotu und Deshpande (2015) nennen den Nachteil, dass die Naive Bayes Klassifikation davon ausgeht, dass die Attribute voneinander unabhängig sind und alle einen gleich großen

Einfluss auf die vorherzusagende Klasse haben. Ein weiterer, bereits zuvor genannter Nachteil ist, dass die Naive Bayes Klassifikation nur mit kategorischen Werten umgehen kann und fortlaufende numerische Werte in solche transformiert werden müssen.

3.6 Neuronale Netzwerke

Kotu und Deshpande (2015) fassen neuronale Netzwerke zur Lösung von Klassifizierungsproblemen als mathematischen Versuch zusammen, welcher das Ziel verfolgt, den Zusammenhang zwischen Eingabeattributen und Ausgabeattributen des Klassifizierungsproblems mathematisch zu beschreiben. Dabei wird der Zusammenhang zwischen Eingangsattributen und Ausgangsattributen in Form eines gewichteten Netzwerkes modelliert, die Attribute stellen dabei Knoten im Netzwerk dar, welche mit gewichteten Verbindungen verbunden werden.

Bishop (2006) beschreibt ein neuronales Netzwerk für Klassifizierungsprobleme als ein Netzwerk, welches aus einer Reihe von Eingabeknoten, welche die Eingabeattribute repräsentieren, sowie aus einer Reihe von Ausgabeknoten, welche die vorherzusagenden Werte repräsentieren, besteht. Diese beiden Arten von Knoten werden durch versteckte Ebenen von Knoten mathematisch miteinander verbunden. Die Verbindungen sind dabei gewichtet. Kotu und Deshpande (2015) ergänzen zu den Eingabeattributen, dass bei neuronalen Netzwerken, ähnlich wie bei linearen Modellen, alle Attribute numerisch sein müssen. Es ist daher bei neuronalen Netzwerken notwendig, klassifizierende Eingabeattribute als eigene Attribute je Klasse abzubilden.

Das Training eines Data Mining Modells mit diesem Data Mining Verfahren erfolgt durch Ermittlung und Anpassung der Gewichtung der Verbindungen im neuronalen Netzwerk. Eines der verbreitetsten Verfahren dafür ist die Fehlerrückführung[6]. Hier werden für jede Instanz im Trainingsdatensatz, die Ausgangsknoten, mit den aktuellen Gewichten berechnet. Wenn die Werte in den Ausgangsknoten den Werten der Ausgangsattribute in der jeweiligen Instanz entsprechen, wurde eine korrekte Vorhersage getroffen. Wenn die Werte jedoch von dem erwarteten abweichen, müssen die Gewichtungen im Netzwerk angepasst werden. Dieses Anpassen der Gewichtungen kann, je nach Implementierung des neuronalen Netzwerkes, unterschiedlich durchgeführt werden. (Han, Kamber, & Pei, 2012)

In Folge wird dieses Data Mining Verfahren anhand der Kriterien nach Kotu und Deshpande (2015) klassifiziert:

Modell: Bei diesem Data Mining Modellen werden die Knoten und deren gewichtete Verbindungen gespeichert. Bei der Vorhersage von neuen Instanzen werden die Eingangsknoten mit den entsprechenden Attributwerten versehen und die Werte der Ausgangsknoten auf Basis der gewichteten Verbindungen errechnet (Bishop, 2006).

[6] Englisch: back propagation

Eingabeattribute: Ähnlich wie beim linearen Modell, müssen hier laut Kotu und Deshpande (2015) alle Eingabeattribute numerisch sein. Sie beschreiben auch, dass klassifizierende Attribute auf ein Eingabeattribut je Klasse abgebildet werden können deren Wert 1 ist, wenn die Klasse vorhanden ist und 0, wenn die Klasse nicht vorhanden ist.

Ausgabeattribute: Han, Kamber und Pei (2012) beschreiben, dass ein neuronales Netzwerk sowohl für Klassifizierungsaufgaben, als auch für die Vorhersage von numerischen Werten eingesetzt werden kann. Bei Vorhersage von zwei Klassen, wird mit einem Ausgabeattribute aus dem Netzwerk gearbeitet und bei Vorhersage von mehreren Klassen, wird je vorherzusagender Klasse ein eigenes Attribut verwendet.

Vorteile: Kotu und Despande (2015) nennen als zwei große Vorteile dieses Data Mining Verfahren die schnelle Verarbeitung von neuen Vorhersagen und sehr gute Vorhersageergebnisse bei nicht linearen Beziehungen in den Daten.

Nachteile Laut Kotu und Deshpande (2015) ist es ein großer Nachteil dieses Data Mining Verfahrens, dass die innere Funktionsweise des Data Mining Modells auf keine einfache Art und Weise visualisiert werden kann. Han, Kamber und Pei (2012) sehen auch den großen Nachteil, dass die Auswahl einer passenden Netzwerktopologie oft ein langwieriger Prozess mit einer Vielzahl an Versuchen und Adaptionen am Netzwerk ist. Neuronale Netzwerke können laut Han, Kamber und Pei (2012) nicht mit fehlenden Werten umgehen.

3.7 Regelinduktion

Beim Data Mining Verfahren der Regelinduktion werden Regeln aus den Daten abgeleitet, anhand deren die Klassifizierung für neue Instanzen durchgeführt werden kann. Die Regeln werden dabei im Format „Wenn-Dann“ gebildet. Wenn eine bestimmte Bedingung zutrifft, dann gehört die Instanz zu einer bestimmten Klasse. Die Bedingungen werden aus Werten und Klassen der bekannten Attribute gebildet, welche mit logischer UND und ODER Verknüpfung zusammengehängt werden können (Han, Kamber, & Pei, 2012). Ein Beispiel für eine sehr einfache solche Regel ist: „Wenn Helligkeit == „Dunkel“, dann Licht=“EIN“. Dieses Beispiel würde für Instanzen bei denen das Attribut Helligkeit der Klasse „Dunkel“ zugewiesen ist, für das Attribut Licht die Klasse „EIN“ vorhersagen.

Beim Erstellen des Data Mining Modells werden die Regeln aus den Daten abgeleitet. Bei der Vorhersage für neue Instanzen werden die Regeln ausgewertet und durch das Zutreffen einer entsprechenden Regel die Klassifizierung der Instanz vorgenommen (Han, Kamber, & Pei, 2012). Laut Kotu und Deshpande (2015) gibt es zwei grundsätzliche Arten, wie die Regeln abgeleitet werden können: Regeln können einfach von einem zuvor aufgebauten Entscheidungsbaum abgeleitet werden, in dem man jeden Pfad im Baum von oben nach unten durchläuft und je Pfad eine Bedingung aus den Knoten erstellt. Dieses Verfahren benötigt als Ausgangsbasis einen Entscheidungsbaum und stellt eigentlich nur ein anderes Format dieses Data Mining Verfahrens dar. Es ist auch möglich, die Regeln direkt aus den Trainingsdaten zu extrahieren. Hierzu werden die Daten Klasse für Klasse durchlaufen und Regeln extrahiert, die für die jeweilige Klasse gelten. Mit jeder Iteration über die Daten wird eine Regel abgeleitet. Aus

diesem Grund spricht man hier auch von sequentieller Regelinduktion. Während der Ableitung einer Regel muss immer überprüft werden, ob die Regel einen Mehrwert für die Vorhersagequalität bietet, bevor sie zum Regelsatz hinzugefügt wird.

In Folge wird dieses Data Mining Verfahren anhand der Kriterien nach Kotu und Deshpande (2015) klassifiziert:

Modell: Das Data Mining Modell bei diesem Data Mining Verfahren stellt eine einfache Liste von Regeln, in Form von „Wenn-Dann“ Bedingungen mit einer bestimmten Reihenfolge dar. Diese Repräsentation ist für Menschen sehr gut verständlich und benötigt keine aufwändigen Algorithmen zur Verwaltung oder zum Lesen. Zur Vorhersage von neuen Instanzen wird diese Liste in der richtigen Reihenfolge durchlaufen. Sobald die Bedingung der ersten Regel zutrifft wird die Instanz durch diese Regel klassifiziert. (Han, Kamber, & Pei, 2012)

Eingabeattribute: Es können sowohl klassifizierende, als auch fortlaufende Attribute als Eingabeattribute verwendet werden (Kotu & Deshpande, 2015).

Ausgabeattribute: Mit diesem Data Mining Verfahren können nur Modelle erstellt werden, welche es ermöglichen, ein klassifizierendes Attribut vorherzusagen (Kotu & Deshpande, 2015).

Vorteile & Nachteile: Durch die nahe Verwandtschaft mit dem Data Mining Verfahren der Entscheidungsbäume ergeben sich hier auch dieselben Vorteile und Nachteile wie bereits in 3.2 beschrieben.

3.8 Lineare Modelle zur Klassifizierung

Witten, Frank, Hall und Pal (2016) beschreiben, dass jede Art der Regression auch für Klassifizierungsaufgaben verwendet werden kann. Man muss für jede Klasse des zu vorherzusehenden Attributes eine Regressionsgleichung aufstellen, in welcher die anderen Attribute des Datensatzes enthalten sind. Für fortlaufende Attributwerte kann der Wert selbst eingesetzt werden. Für klassifizierende Attribute muss für jede mögliche Klasse, ein eigenes Attribut in die Gleichung aufgenommen werden. Wenn ein Attribut einer bestimmten Klasse angehört, ist der jeweilige Wert 1 und alle anderen Klassen dieses Attributes haben den Wert 0. So wird eine jede Klasse numerisch auf 1 und 0 abgebildet. 1 wenn die Klasse vorhanden ist und 0 wenn die Klasse nicht vorhanden ist. Auch die vorherzusagende Klasse wird auf diesen Nummernbereich zwischen 0 und 1 abgebildet. Bei der Vorhersage wird die Regressionsgleichung für alle Klassen des vorherzusagenden Attributes aufgestellt und gelöst. Die Klasse mit dem höchsten Wert ist die vorhergesagte Klasse (Witten, Frank, Hall, & Pal, 2016).

Bishop (2006) beschreibt, dass die logistische Regression als Data Mining Modell für Klassifizierungsaufgaben gut geeignet ist. Die logistische Regression ist für Zwei-Klassen Probleme geeignet. Witten, Frank, Hall und Pal (2016) ergänzen hier, dass mit der logistischen Regression statistische Probleme anderer Regressionsmethoden für diesen Anwendungsfall vermieden werden können.

In Folge wird dieses Data Mining Verfahren anhand der Kriterien nach Kotu und Deshpande (2015) klassifiziert:

Modell: Bei diesen Data Mining Modellen handelt es sich um Regressionsgeraden, welche mit den Koeffizienten, je Attribut, ausgestattet sind. Die Vorhersage von neuen Instanzen ist sehr einfach und effizient durch Einsetzen der bekannten Werte in die Gleichung und Lösung dieser möglich (Kotu & Deshpande, 2015).

Eingabeattribute: Laut Kotu und Deshpande (2015) müssen alle Eingabeattribute numerisch sein. Witten, Frank, Hall und Pal (2016) beschreiben, dass klassifizierende Attribute für die Regressionsgleichung auf ein eigenes Attribut je Klasse abgebildet werden können. Wenn deren Wert 1 ist, ist die Klasse vorhanden, wenn deren Wert 0 ist, ist die Klasse nicht vorhanden.

Ausgabeattribute: Regressionen sind generell für fortlaufende Attributstypen gedacht, es ist jedoch auch möglich, Klassifizierungsaufgaben mit Regression zu lösen, in dem eine logistische Regression verwendet wird oder die Klassen auf die Zahl 1, wenn vorhanden und 0, wenn nicht vorhanden, abgebildet werden. Mit einer logistischen Regression kann nur ein Zwei-Klassen Problem gelöst werden. Es gibt aber auch Ansätze um mehr als zwei Klassen mit diesem Data Mining Verfahren vorhersagen zu können (Bishop, 2006).

Vorteile: Dieses Data Mining Verfahren ist weit verbreitet und aus Sicht der Rechenleistung effizient. Neue Vorhersagen mit einem fertigen Modell können durch einfaches Lösen von Geradengleichungen durchgeführt werden (Kotu & Deshpande, 2015).

Nachteile: Kotu und Deshpande (2015) nennen als großen Nachteil, dass dieses Data Mining Verfahren nicht mit fehlenden Daten umgehen kann. Außerdem wird die Repräsentation des Data Mining Modells sehr unübersichtlich wenn eine große Anzahl von Attributen und Klassen beteiligt ist.

3.9 Anforderungen an den Anwendungsfall

In diesem Abschnitt wird auf die Anforderungen an ein Data Mining Verfahren eingegangen, welches für den Anwendungsfall in der vorliegenden Arbeit eingesetzt wird. Diese Anforderungen bilden die Basis für die Auswahl eines geeigneten Data Mining Verfahrens für die Umsetzung des Prototyps. Bei der Definition und Auswahl der Anforderungen wurde die zuvor getroffene Klassifizierung der Data Mining Verfahren, sowie der bereits zuvor beschriebene CRISP-DM Prozess berücksichtigt, um relevante Anforderungen identifizieren zu können.

Wie in den Arbeiten von Cook et al. (2003) und Das, Chen, Seelye und Cook (2011) beschrieben, werden Daten in einem Smart Home anhand von Events, welche von den installierten Sensoren generiert werden, gesammelt. Die vernetzten Sensoren und Aktoren in einem Smart Home generieren Events, wenn sich ein Zustand oder ein Wert ändert. Laut Das, Chen, Seelye und Cook (2011) besteht ein Event aus den Informationen: Zeitpunkt, Sensor welcher das Event gesendet hat, sowie der Nachricht die gesendet wurde. Diese Events

werden in einem Smart Home, wie es in der vorliegenden Arbeit im Einsatz ist, zentral gesammelt und in einer Datenbank mit Zeitbezug persistiert. Die Nachricht kann dabei rein numerische Messwerte, aber auch klassifizierende Attribute beinhalten. Als Beispiele für rein numerische Messwerte nennen Das, Chen, Seelye und Cook (2011) Temperatursensoren und den aktuellen Energieverbrauch. Als Beispiele für klassifizierende Attribute, nennen sie Bewegungsmelder (Bewegung erkannt und keine Bewegung mehr erkannt) oder den Zustand einer Lichtinstallation (Licht ein und Licht aus). Für die Auswahl eines Data Mining Verfahrens für die vorliegende Arbeit ergibt sich daraus die Anforderung, dass das Data Mining Verfahren sowohl mit fortlaufenden numerischen Attributen, als auch mit klassifizierenden Attributen trainiert werden muss.

Durch die eventbasierte Datensammlung der einzelnen Attribute kann zu jedem beliebigen Zeitpunkt innerhalb der Aufzeichnungen, der zu diesem Zeitpunkt aktive Zustand oder Wert eines Sensors oder Aktors, nachvollzogen werden. Durch diese Art der Datensammlung kommt es zu keinen unvollständigen Instanzen. In Zeiträumen in denen zum Beispiel ein Sensor defekt ist und daher keine Events generiert, zählt der Wert des letzten Events vor dem Defekt. Somit ergibt sich für die Auswahl eines Data Mining Verfahrens, für die vorliegende Arbeit, keine Anforderung, dass das Data Mining Modell mit fehlenden Attributwerten umgehen muss. Für einen Einsatz in der Praxis wird eine Lösung für diese Problematik gefunden werden müssen, da sich hier auch die Zahl der Sensoren durch Änderungen am Smart Home verändern kann und sich dadurch unvollständige Instanzen über gewisse Zeiträume ergeben. Dieses Thema wird in der vorliegenden Arbeit nicht berücksichtigt, da keine Veränderungen am Smart Home im Evaluierungszeitraum vorgenommen werden.

Im Anwendungsfall der vorliegenden Arbeit soll der Zustand von Lichtinstallationen vorausgesagt werden, wenn ein Bewohner oder eine Bewohnerin einen Raum betritt oder verlässt. Die Lichtinstallationen sind dabei als einfache Beleuchtungen definiert, welche die zwei Zustände „Licht An“ und „Licht Aus“ kennen. Für die Auswahl eines Data Mining Verfahrens für die vorliegende Arbeit ergibt sich daraus die Anforderung, dass das Data Mining Verfahren Vorhersagen zu einem klassifizierenden Attribut mit zwei möglichen Klassen treffen muss.

Die Vorhersagen sollen direkt beim Betreten oder Verlassen eines Raumes durch einen Bewohner oder eine Bewohnerin getroffen werden. Im Optimalfall soll der Bewohner oder die Bewohnerin keine spürbare Verzögerung zwischen dem Betreten des Raumes und der eventuell folgenden Aktion (Licht wird ein- oder ausgeschalten) bemerken. Für die Auswahl eines Data Mining Modells, für die vorliegende Arbeit, ergibt sich daraus die Anforderung, dass die Vorhersagegeschwindigkeit schnell sein muss.

Wie bereits in 2.3.3 beschrieben, fallen in einem Smart Home, wie es in der vorliegenden Arbeit verwendet wird, circa ein Gigabyte neue Daten pro Jahr an. Nicht alle erfassten Attribute und Events werden für die Anwendung im Data Mining Modell der vorliegenden Arbeit benötigt. Das Training des Data Mining Modells für die vorliegende Arbeit muss regelmäßig neu stattfinden können. Daraus ergibt sich für die Auswahl eines Data Mining Verfahrens die Anforderung, dass mit dem gewählten Data Mining Verfahren mit Trainingsdaten im einstelligen Gigabyte

Bereich das Data Mining Modell während der Evaluierungsphase mehrmals neu aufgebaut werden kann.

Zusammengefasst wurden folgende Anforderungen für die Auswahl eines Data Mining Verfahrens für die vorliegende Arbeit identifiziert:

- Das gewählte Data Mining Verfahren muss in der Lage sein, sowohl mit numerischen fortlaufenden Attributen, als auch mit klassifizierenden Attributen umzugehen.
- Das gewählte Data Mining Verfahren muss in der Lage sein, ein Attribut mit zwei Klassen vorauszusagen.
- Die Vorhersage einer neuen Instanz mit dem erstellten Data Mining Modell, muss so schnell möglich sein, dass für einen Menschen keine spürbare Verzögerung entsteht.
- Das gewählte Data Mining Verfahren muss in der Lage sein, mit Trainingsdaten im einstelligen Gigabytebereich umgehen zu können.
- Das Data Mining Modell muss in, für die Evaluierungsphase der vorliegenden Arbeit, akzeptabler Zeit neu aufgebaut werden können. So dass während dem Evaluierungszeitraum ein mehrfaches Aktualisieren des Data Mining Modells möglich ist.

Diese Anforderungen bilden die Basis für die Auswahl eines Data Mining Verfahrens für die Umsetzung des Prototyps in der vorliegenden Arbeit.

3.10 Auswahl eines Data Mining Verfahrens für den Prototyp

Ziel dieses Abschnittes ist es, eine Auswahl aus den, in Abschnitten 3.2 bis 3.8, analysierten Data Mining Verfahren für die Umsetzung des Prototyps in der vorliegenden Arbeit zu treffen. Dabei wurde das Data Mining Verfahren ausgewählt, welches die größte Genauigkeit bei den Vorhersagen für den Anwendungsfall in der vorliegenden Arbeit erzielen konnte und welches alle Anforderungen an den Anwendungsfall erfüllt.

Das „No-Free-Lunch-Theorem“ von Wolpert und Macready (1995), welches von Flach (2012) auf Klassifizierungsprobleme angewendet wurde, besagt, dass kein Data Mining Verfahren für Klassifizierungsaufgaben generell als besser, als andere Data Mining Verfahren bewertet werden kann, wenn man alle Klassifizierungsprobleme die existieren betrachtet. In dieser Betrachtung ist die Genauigkeit keines der Data Mining Verfahren besser, als einfaches Erraten der Vorhersagen. Was Flach damit beschreiben will, ist, dass es im Bereich des Data Mining keine perfekte Lösung für ein Problem gibt. Es ist daher notwendig, dass man für ein spezifisches Klassifizierungsproblem die Problemstellung und die Ergebnisse einzelner Data Mining Verfahren im Detail analysiert und anschließend jenes Data Mining Verfahren auswählt, welches für diesen Fall die beste Lösung darstellt. Für Vergleiche von Data Mining Verfahren wurden im Laufe der letzten Dekaden eine große Zahl von statistischen Verfahren entwickelt. Demsar (2006) hat Publikationen über den Vergleich von Data Mining Verfahren analysiert und die eingesetzten Werkzeuge verglichen. Er hat zusammengefasst, dass die Kreuzvalidierung

für diese Art von Vergleichen, die am häufigsten eingesetzte Evaluierungsmethode ist. Für die Bewertung des Data Mining Verfahren selbst, sind die Attribute der Genauigkeit, der Präzision und der Trefferquote als weit verbreitete Attribute identifiziert worden. Zusätzlich werden in vielen Publikationen auch statistische Tests, wie der t-Test oder der 5x2cv Test von Dietterich (1998) durchgeführt. Für die vorliegende Arbeit sollen die oben identifizierten Attribute, sowie entsprechende t-Tests, die Basis für den Vergleich bilden.

Die beiden Anforderungen betreffend der Eingabe-, sowie der Ausgabeattribute konnten durch die getroffene Klassifizierung der Data Mining Verfahren geprüft werden. Die Anforderungen betreffend der Datenmenge, der Vorhersage- und Trainingsgeschwindigkeit, sowie der Genauigkeit der Vorhersage, konnten nur durch einen Vergleich anhand eines trainierten Data Mining Modells geprüft werden. In Tabelle 2 ist die Gegenüberstellung der Abdeckung der Anforderungen an die Eingabe- und Ausgabeattribute dargestellt. Darin ist ersichtlich, dass die Data Mining Verfahren Entscheidungsbäume und Regelinduktion die Anforderungen komplett abdecken. Die anderen fünf Data Mining Verfahren erwarten rein numerische Eingabeattribute, es gibt jedoch für alle fünf Data Mining Verfahren Ansätze, um auch klassifizierende Attribute als Eingabeattribute verwenden zu können. Wie in 3.3 und 3.4 beschrieben, ist es möglich, bei den Data Mining Verfahren der Stützvektormaschinen, sowie bei der „Nächster Nachbar Klassifizierung“, die klassifizierenden Attribute auf einen fortlaufenden numerischen Bereich abzubilden. Diese Abbildung bringt hier aber unerwünschte Seiteneffekte durch die dadurch entstehende Reihenfolge der Klassen. Bei den Data Mining Verfahren neuronale Netzwerke und lineare Modelle können klassifizierende Attribute, wie in den Abschnitten 3.6 und 3.8 beschrieben, durch eigene Eingabeattribute je möglicher Klasse abgebildet werden. Diese Vorgehensweise ist hier auch sehr verbreitet und beliebt. Bei der Naive Bayes Klassifizierung, können, wie in 3.5 beschrieben, nur klassifizierende Eingabeattribute verwendet werden. Hier gibt es die Möglichkeit fortlaufende numerische Werte auf Wertintervalle abzubilden, bei denen jeder Intervallbereich eine Klasse repräsentiert. Zusammengefasst erfüllen die fünf Data Mining Verfahren Entscheidungsbäume, Regelinduktion, neuronale Netzwerke, lineare Modelle und die Naive Bayes Klassifizierung die Anforderungen an die Eingabe- und Ausgabeattribute an den Anwendungsfall in der vorliegenden Arbeit. Um Seiteneffekte durch die Umschlüsselung der klassifizierenden Attribute zu vermeiden, werden die beiden Data Mining Verfahren der Stützvektormaschinen und der „nächster Nachbar Klassifizierung“ in der vorliegenden Arbeit nicht weiter berücksichtigt.

<table>
<tr><th>Data Mining Verfahren</th><th>Anforderung an Eingabeattribute</th><th>Anforderung an Ausgabeattribute</th></tr>
<tr><td>Entscheidungsbäume</td><td>Vollständig abgedeckt</td><td>Vollständig abgedeckt</td></tr>
<tr><td>Regelinduktion</td><td>Vollständig abgedeckt</td><td>Vollständig abgedeckt</td></tr>
<tr><td>Stützvektormaschinen</td><td rowspan="3">Eingabeattribute müssen numerisch sein. Die Abbildung von klassifizierenden Attributen als numerische Attribute ist</td><td rowspan="3">Vollständig abgedeckt</td></tr>
<tr><td>Nächster Nachbar Klassifizierung</td></tr>
<tr><td>Neuronale Netzwerke</td></tr>
</table>

Lineare Modelle	jedoch möglich.	
Naive Bayes Klassifizierung	Eingabeattribute müssen klassifizierend sein. Die Abbildung von numerischen Attributen als klassifizierende Attribute ist jedoch möglich.	Vollständig abgedeckt

Tabelle 2 Abdeckung der Anforderung an Eingabe- sowie Ausgabeattribute

Um die Erfüllung der drei weiteren Anforderungen, sowie die Genauigkeit der einzelnen Data Mining Verfahren, für den Anwendungsfall der vorliegenden Arbeit, ermitteln zu können, war es notwendig die fünf Data Mining Verfahren anhand eines Beispiels miteinander zu vergleichen. Für den Vergleich wurde ein Datensatz, welcher einen Zeitraum von 30 Tagen abdeckt, mit den in den Abschnitten 4.2.1, 4.2.2 und 4.2.3, durch den CRISP-DM Prozess, definierten Attribute und Struktur ausgewählt. Dieser Datensatz wurde verwendet, um für jedes der ausgewählten Data Mining Verfahren, ein Data Mining Modell für eine ausgewählte Beleuchtung zu trainieren. Die Ergebnisse der Modelle wurden anschließend verglichen, um das beste Data Mining Verfahren für den Anwendungsfall der vorliegenden Arbeit und somit für den Bau des Prototyps auszuwählen. Für diesen Vergleich wurde, die für Experimente und Versuche gedachte, freie Data Mining Software Weka[7] eingesetzt. Diese Software von der Universität Waikato ist dazu gedacht, mit Data Mining Verfahren einfach und schnell, Experimente und Evaluierungen durchführen zu können. Die Software unterstützt die gängigsten Data Mining Verfahren mit unterschiedlichen Implementierungen. Für die Evaluierung des Data Mining Verfahrens Entscheidungsbäume, wurde der weit verbreitete Algorithmus J48 ausgewählt. Dieser ist eine Java Implementierung des weit verbreiteten C4.5 Algorithmus. Für die Evaluierung der Regelinduktion wurde auf den Algorithmus JRip zurückgegriffen. Für die Evaluierung der Umsetzung mit einem neuronalen Netzwerk wurde auf ein einfaches neuronales Netzwerk nach dem Prinzip der Fehlerrückführung zurückgegriffen, dabei wurde die Anzahl der benötigten versteckten Schichten im Netzwerk durch Weka automatisch ermittelt. Für die Evaluierung der Vorhersage mit linearen Modellen wurde auf eine einfache logistische Regression zurückgegriffen, da diese für die Vorhersage von einem zwei Klassenproblem als am besten geeignet identifiziert wurde. Für alle Data Mining Verfahren wurde die von Weka vorgeschlagenen Standardparametrisierung der Algorithmen verwendet. Es wurde keine Optimierung der Parameter durchgeführt.

Mit den ausgewählten Data Mining Verfahren und dem gebildeten Datensatz, wurde eine 5x2cv Validierung durchgeführt. Dabei wurde eine Kreuzvalidierung mit zwei Segmenten fünf Mal wiederholt. Mit denselben Segmenten wurde ein Data Mining Modell mit jedem der ausgewählten Verfahren trainiert und validiert. Anschließend wurden die gesammelten Messwerte der Iterationen und der Unterschiede zwischen den einzelnen Data Mining Verfahren mit paarweisen t-Tests analysiert. Die gesammelten Ergebnisse für die analysierten

[7] http://www.cs.waikato.ac.nz/ml/weka/

Data Mining Verfahren sind in Tabelle 3 ersichtlich. Dort ist ersichtlich, dass das Data Mining Verfahren der linearen Modelle, sowie die Naive Bayes Klassifiktation, für den Anwendungsfall der vorliegenden Arbeit deutlich schlechtere Vorhersagen trafen, als die anderen drei untersuchten Data Mining Verfahren. Die besten Leistungen erbrachten in den Versuchen, die Data Mining Verfahren der Entscheidungsbäume und der Regelinduktion. Beide erzielten eine sehr ähnliche Vorhersagequalität. Die neuronalen Netzwerke erzielten eine etwas schlechtere, aber dennoch sehr gute Vorhersagequalität für den untersuchten Datensatz. Die Data Mining Verfahren der linearen Modelle und der Naive Bayes Klassifikation werden für die vorliegende Arbeit und für die weitere Auswahl eines Data Mining Verfahrens, für den Bau des Prototyps, nicht weiter berücksichtigt.

Data Mining Verfahren	True Positive	False Positive	True Negative	False Negativ	Präzision	Trefferquote
Entscheidungsbäume	99,74%	0,15%	99,84%	0,25%	99,77%	99,75%
Regelinduktion	99,78%	0,12%	99,83%	0,21%	99,83%	99,79%
Neuronale Netzwerke	99,11%	0,74%	99,26%	0,89%	98,93%	99,11%
Lineare Modelle	94,63%	3,8%	96,19%	5,36%	94,47%	94,634%
Naive Bayes	97,08%	9,6%	90,35%	2,9%	87,34%	97,081%

Tabelle 3: Vergleich der ausgewählten Data Mining Verfahren

Das Training der Data Mining Modelle war für alle Data Mining Verfahren in wenigen Minuten möglich und die Vorhersage für neue Instanzen war bei allen Data Mining Verfahren in wenigen Millisekunden möglich. Daher konnte auf Basis der beiden Anforderungen bezüglich Trainingszeit und Vorhersagezeit kein Data Mining Verfahren für den Anwendungsfall der vorliegenden Arbeit ausgeschlossen werden.

Der gepaarte t-Test über die drei verbliebenen Data Mining Verfahren Entscheidungsbäume, Regelinduktion und neuronale Netzwerke hat zwischen Entscheidungsbäumen und neuronalen Netzwerken, sowie der Regelinduktion und neuronalen Netzwerken einen signifikanten Unterschied ergeben. In beiden Fällen wurden die neuronalen Netzwerke in der Betrachtung der Präzision und der Trefferquote als signifikant schlechter eingestuft. Zwischen den beiden Data Mining Modellen Entscheidungsbäume und Regelinduktion hingegen konnte kein signifikanter Unterschied identifiziert werden. Von den beiden verbleibenden Data Mining Verfahren wurde das Data Mining Verfahren der Entscheidungsbäume für die Umsetzung des Prototyps in der vorliegenden Arbeit ausgewählt, da dieses Data Mining Verfahren sehr weit verbreitet ist und sich auch in den meist verbreitetsten Data Mining Verfahren laut Wu et al. (2007) mehrfach wiederfindet.

3.11 Zusammenfassung

In diesem Kapitel wurde zuerst eine Literaturrecherche in Fachbüchern zu Data Mining Verfahren für Klassifizierungen durchgeführt. Die Ergebnisse dieser Literaturrecherche wurden

aufgelistet und auf die Auflistung eingegrenzt. Die Data Mining Verfahren Entscheidungsbäume, Stützvektormaschinen, „Nächster Nachbarn Klassifikation“, „Naive Bayes Klassifikation„, „Neuronale Netzwerke“, Regelinduktion und „Lineare Modelle“, wurden auf Basis dieser Literaturrecherche ausgewählt und in Folge näher betrachtet.

Die Funktionsweise eines jeden dieser ausgewählten Data Mining Verfahren wurde erörtert und zusammengefasst. Auf dieser Basis wurde eine Klassifizierung der Data Mining Verfahren durchgeführt. Mit dieser Klassifizierung, nach Kotu und Deshpande (2015), wurde das Ziel verfolgt die Auswahl eines Data Mining Verfahrens für den Prototyp der vorliegenden Arbeit zu ermöglichen. Die Klassifizierung wurde nach der Art des entstehenden Modells, der Attributtypen, welche als Eingabeattribute und als Vorhersageattribute verwendet werden können, sowie nach in der Literatur gefundenen Vor- und Nachteilen der jeweiligen Data Mining Verfahren durchgeführt.

Nachdem eine Auflistung und Klassifizierung der Data Mining Verfahren für den Anwendungsfall der vorliegenden Arbeit aufgestellt war, wurden die Anforderungen an diese Data Mining Modelle für den Anwendungsfall analysiert und definiert. Dabei wurden die in einem Smart Home auftretenden Datenstrukturen und Datentypen analysiert und festgestellt, dass das Data Mining Verfahren sowohl mit fortlaufenden, als auch klassifizierenden Attributen umgehen muss. Für die Vorhersage wurde identifiziert, dass ein klassifizierendes Attribut mit zwei möglichen Klassenwerten vorhergesagt werden muss. Es wurden auch Anforderungen an die Trainings- und Vorhersagegeschwindigkeit aufgestellt.

Auf Basis dieser Anforderungen konnte die Auswahl der passenden Data Mining Verfahren bereits eingegrenzt werden. Die Data Mining Verfahren der Stützvektormaschinen und der „nächsten Nachbarn Klassifikation“ wurden auf Grund der Anforderungen ausgeschlossen. Für die Auswahl des, für die vorliegende Arbeit am besten geeigneten Data Mining Verfahrens, wurde im Anschluss mit einem Trainingsdatensatz für jedes der Data Mining Verfahren ein Data Mining Modell trainiert und mit der 5x2cv Validierung evaluiert. Im Zuge dieser Validierung konnten die Data Mining Verfahren der „neuronalen Netzwerke“, der „lineare Modelle“ und der „Naive Bayes Klassifikation“ für die vorliegende Arbeit ausgeschlossen werden. Die beiden Data Mining Verfahren Entscheidungsbäume und Regelinduktion konnten in diesem Vergleich beide sehr gut abschneiden. Abschließend wurde das Data Mining Verfahren der Entscheidungsbäume für die Verwendung im Prototyp der vorliegenden Arbeit ausgewählt. Auf Basis dieser Auswahl wird, im folgenden Kapitel, der CRISP-DM Prozess durchlaufen und der Aufbau des Prototyps für die Evaluierung beschrieben.

4 AUFBAU DES DATA MINING MODELLS UND PROTOYPS

In diesem Kapitel wird zuerst, auf Basis des zuvor in Kapitel 3 ausgewählten Data Mining Verfahrens der Entscheidungsbäume und der bestehenden Smart Home Infrastruktur, eine Auswahl der für die Implementierung des Prototyps notwendigen Technologien getroffen. Anschließend wird auf Basis der Technologieauswahl und des ausgewählten Data Mining Verfahrens der CRISP-DM Prozess durchlaufen, um ein Data Mining Modell für den Prototyp der vorliegenden Arbeit zu erstellen. Anschließend werden die Struktur, der Aufbau und die Prozesse des Prototyps beschrieben, welcher in das Smart Home zur Evaluierung und zur Beantwortung der Forschungsfrage integriert wird. Abschließend wird auf die implementierten Prozesse und Funktionen des Prototyps für die Evaluierung näher eingegangen.

4.1 Auswahl der Technologien

Die Smart Home Software, welche in dem Smart Home, welches für die Evaluierung verwendet wird, im Einsatz ist, ist OpenHab[8]. Die in Java geschriebene Open Source Smart Home Software OpenHab findet aktuell große Beliebtheit und große Verbreitung. Durch ein integriertes Modulsystem ist es möglich, auf einfache Art und Weise neue Komponenten oder andere System in die zentrale Smart Home Lösung zu integrieren. Es existieren bereits eine Vielzahl von sogenannten Bindings, mit denen eine sehr große Anzahl von Smart Home Komponenten verschiedenster Hersteller und damit verwandte Systemen, in das Smart Home integriert werden können. Die OpenHab Installation, im Smart Home der vorliegenden Arbeit, persistiert alle gesammelten Daten in zwei Datenbanksystemen: MariaDB[9] und InfluxDB[10]. MariaDB ist eine klassische relationale Datenbank, welche aus dem MySQL Projekt hervorgegangen ist. InfluxDB ist ein relativ neues Datenbanksystem, welches speziell auf die Verwaltung von Zeitreihen ausgerichtet ist. Beide Datenbanksysteme sind in die zentrale OpenHab Software mit entsprechenden Bindings eingebunden. Genauere Informationen zur Anbindung und Verwendung dieser Datenbanksysteme wurden in Abschnitt 4.2.3 analysiert. Für die Anbindung von Komponenten im Smart Home, welche über kein eigenes OpenHab Binding verfügen, wird das Bussystem MQTT[11] eingesetzt. Dieser Enterprise Service Bus ist speziell für die Maschine zu Maschine Kommunikation im Internet der Dinge und damit auch im Smart Home gedacht. Konkret ist die Open Source Implementierung Mosquitto[12] im Einsatz.

[8] https://www.openhab.org/

[9] https://mariadb.org/

[10] https://www.influxdata.com/time-series-platform/influxdb/

[11] http://mqtt.org/

[12] https://mosquitto.org/

Dieses Bussystem ist mit einem Binding in OpenHab integriert. OpenHab ist somit in der Lage Nachrichten über dieses Bussystem zu erhalten oder zu versenden. Im gesamten Smart Home ist dieses Bussystem, sowie die beiden Datenbank Instanzen über das lokale Netzwerk erreichbar. Aus Infrastruktursicht wird ein jedes, der oben beschriebenen Services, in eigenen Docker[13] Containern, auf einer lokalen Containerinfrastruktur im Smart Home, betrieben. Der Prototyp, welcher in der vorliegenden Arbeit entwickelt wurde, musste in diese Infrastruktur Umgebung integriert werden.

Für die Umsetzung der Kernkomponenten des Prototyps, den Aufbau und die Evaluierung des Data Mining Modells, sowie das Treffen von neuen Vorhersagen, wurden die Open Source Datenverarbeitungsframeworks Apache Spark[14] und das bereits für die Auswahl des Data Mining Verfahrens verwendete Programm, Weka untersucht. Dabei wurde festgestellt, dass Weka für den Bau eines Prototyps für die vorliegende Arbeit, wesentlich besser geeignet ist, als Apache Spark. Apache Spark ist auf sehr große Datenmengen und stark verteiltes Data Mining, auf mehreren Rechnerknoten, ausgelegt. Der Umgang mit dem Framework für einen einfachen Prototyp ist damit wesentlich komplexer und aufwändiger als bei Weka. Weka bietet eine Vielzahl an Implementierungen für das Data Mining Verfahren der Entscheidungsbäume, welches in Abschnitt 3.10 für die Umsetzung des Prototyps ausgewählt wurde. Die Weka Bibliothek bietet eine einfache API (Applikation Programming Interface) an, über welche mit Java einfach Data Mining Modelle trainiert und evaluiert werden können. Für die Integration des Weka Data Mining Modells in das Smart Home, wurde das Java Framework Spring Boot[15] als Basisframework gewählt. Für Spring Boot gibt es für die benötigten Datenbanken, sowie MQTT und für Weka geeignete Bibliotheken, welche direkt durch Konfiguration integriert werden können. Spring Boot Apps können in einem Docker Container betrieben werden und im Internet gibt es eine große Anzahl an Artikeln und Anleitungen für den Umgang mit den benötigten Spring Komponenten.

Für die Detektion des Ereignisses, wenn ein Bewohner oder eine Bewohnerin einen Raum betritt, wurde auf bereits vorhandene, aber noch nicht im Einsatz befindende, Bluetooth Beacons mit Googles Eddystone[16] Protokoll, in Kombination mit einer neuen Android Smartphone App, zurückgegriffen. Bluetooth Beacons sind kleine batteriebetriebene Geräte, welche in zyklischen Abständen über Bluetooth ein Signal aussenden. Dieses Signal kann von einem Smartphone empfangen und ausgewertet werden. Auf Basis des Signals und der Signalstärke kann die Entfernung zum Beacon errechnet werden. Durch diese Auswertung ist es möglich zu identifizieren, wenn ein Bewohner oder eine Bewohnerin mit entsprechendem Smartphone einen Raum betritt oder verlässt. Für diese Auswertung und der Kommunikation der identifizierten Ereignisse an das Smart Home wird, auf eine neue Android App inklusive MQTT Anbindung zurückgegriffen. Wenn die App ein neues Ereignis erkennt, wird dieses über

[13] https://www.docker.com/

[14] https://spark.apache.org

[15] https://projects.spring.io/spring-boot/

[16] https://developers.google.com/beacons/

den MQTT Bus an das Smart Home übermittelt, damit eine entsprechende Vorhersage über das Data Mining Modell durchgeführt werden kann.

Zusammengefasst wurden als Technologien für die Implementierung des Data Mining Modells Java, Spring Boot und die Weka Data Mining Bibliothek ausgewählt. Weka wird für das Trainieren, Evaluierung und Ausführen des Data Mining Modells verwendet. Eine Spring Boot Applikation wird für Integration in das Smart Home, die Prozesssteuerung, sowie für die Datenextraktion und Aufbereitung verwendet. Ereignisse für das Auslösen einer neuen Vorhersage werden mit einer Android App auf den Smartphones der Bewohner und Bewohnerinnen erkannt, welches das Betreten eines Raumes durch die Näherung an eine Bluetooth Beacon erkennt und dieses Ereignis über den vorhandenen MQTT Bus an die Spring Boot Applikation übermittelt. Diese löst auf Basis dieses Ereignisses eine Vorhersage im Data Mining Modell aus und leitet gegebenenfalls eine Statusänderung der Beleuchtung über den MQTT Bus an die Smart Home Applikation OpenHab weiter. Für die Integration in das bestehende Smart Home wird auf den bestehenden MQTT Bus zurückgegriffen. Die Datenbeschaffung für das Trainieren des Data Mining Modells erfolgt durch direkten Zugriff der Applikation auf die eingesetzten Datenbanksysteme. Für die Aufzeichnung aller Aktivitäten im Prototyp, sowie für die Verwaltung der trainierten Data Mining Modelle, wird auf eine eigene Datenbank am bestehenden MariaDB Server zurückgegriffen.

4.2 Entwurf und Implementierung des Data Mining Modells

In diesem Abschnitt werden die Ergebnisse der durchlaufenen Prozessschritte, fachliches Verständnis, Verständnis der Daten, Datenaufbereitung, Modellierung, Evaluierung und Verwendung des, in Abschnitt 2.2 beschriebenen, CRISP-DM Prozesses für den Anwendungsfall der vorliegenden Arbeit beschrieben.

4.2.1 Fachliches Verständnis

Wie in Abschnitt 2.2 beschrieben, besteht die Hauptaufgabe dieses Prozessschrittes darin, eine Anforderungsanalyse für das Data Mining Problem durchzuführen. Ziel des Prototyps der vorliegenden Arbeit ist es mittels Data Mining vorauszusagen, wann eine Beleuchtung im Smart Home eingeschalten oder ausgeschalten werden muss. Um diese Entscheidung treffen zu können, ist es notwendig zu wissen, unter welchen Bedingungen und bei welchen Ereignissen eine Beleuchtung aktiv sein soll und unter welchen Bedingungen keine Beleuchtung benötigt wird. Um die relevanten Bedingungen und Ereignisse besser beurteilen zu können, wurde eine Literaturrecherche in Fachzeitschriften aus dem Bereich Gebäudesimulation nach Artikeln über den Einsatz von künstlicher Beleuchtung durchgeführt. Drei relevante Ergebnisse dieser Recherche werden in Folge kurz beschrieben. Als relevant wurden Arbeiten eingestuft, welche sich mit der Beleuchtungsnutzung in Büro oder Wohnungen, im Zusammenhang mit der Anwesenheit von Personen und der Aktivität im jeweiligen Raum, auseinandergesetzt haben.

Mahdavi, Mohammadi, Kabir und Lambeva (2008) haben in einer Studie untersucht, ob sich das Verhalten zur Beleuchtungsaktivierung in Büros aus Umgebungsattributen ableiten lässt und sind dabei zum Schluss gekommen, dass sich das Verhalten zur Nutzung von Lichtinstallation von Umgebungsattributen ableiten lässt. Dabei haben sie die Attribute Raumtemperatur, Außentemperatur, Luftfeuchtigkeit, Helligkeit, Status der Lichtinstallation, Status der Beschattung, Sonneneinstrahlung, Windgeschwindigkeit und die Präsenz der Menschen im Büro aufgezeichnet und ausgewertet.

Love (1998) beschreibt, dass es grundsätzlich zwei Verhaltensmuster bei der Beleuchtungsaktivierung gibt. Beim ersten Verhaltensmuster wird die Lichtinstallation zu Beginn einer Aktivität eingeschaltet und erst nach Beenden der Aktivität wieder ausgeschaltet. Beim zweiten Verhaltensmuster wird die Lichtinstallation aktiviert, sobald der Helligkeitswert im Raum als zu gering wahrgenommen wird.

Hunt (1979) stellte eine Korrelation zwischen der Helligkeit im Raum und der Wahrscheinlichkeit, dass eine Beleuchtung aktiviert wird, her. Je niedriger die Helligkeit im Raum, desto wahrscheinlicher, dass die Beleuchtung durch einen Bewohner oder eine Bewohnerin aktiviert wird. Zusätzlich stellte er einen Bezug der Beleuchtungsaktivierung und dem Betreten eines Raumes her. Eine Beleuchtung wird meist aktiviert, wenn beim Betreten des Raumes die Helligkeit nicht ausreichend ist. Wenn die Helligkeit im Raum sinkt, während jemand in diesem Raum ist, ist die Wahrscheinlichkeit geringer, dass die Beleuchtung aktiviert wird.

Aus dieser Literaturrecherche ließen sich einige fachliche Anforderungen an das Data Mining Modell ableiten. Es konnten sowohl Anforderungen für den Entwurf und die Planung des Prototypen, sowie für die Erstellung des Vorhersage- und Trainingsprozess in diesem, erhoben werden. Zusätzlich konnten auch Erkenntnisse zu den Daten, die dafür in Betracht gezogen werden müssen, gewonnen werden. Die erhobenen Anforderungen und Erkenntnisse zur Beleuchtungsnutzung sind in Folge beschrieben.

Da laut Hunt (1979) beim Betreten eines Raumes die Wahrscheinlichkeit am höchsten ist, dass eine Beleuchtung eingeschalten werden muss, wurde für den Prototypen der vorliegenden Arbeit definiert, dass die Vorhersage des Status einer Beleuchtung, durch das Betreten eines Raumes, durch einen Bewohner oder eine Bewohnerin, ausgelöst werden muss. Umgekehrt dazu muss auch beim Verlassen des Raumes durch einen Bewohner oder eine Bewohnerin der Status der Beleuchtung, durch eine neuerliche Vorhersage geprüft werden. Ziel dieser Prüfung ist festzustellen, ob eine eingeschaltete Beleuchtung wieder ausgeschaltet werden soll oder ob sie eingeschaltet bleiben soll.

Die Ergebnisse von Mahdavi, Mohammadi, Kabir und Lambeva (2008), sowie auch von Love (1998) und Hunt (1979) lassen darauf schließen, dass der Zustand einer Beleuchtung auf Basis von historischen Aufzeichnungen des Zustandes der Lichtinstallation selbst, in Kombination mit anderen, im Smart Home gesammelten, Umgebungsattributen abgeleitet und vorhergesagt werden kann. Bei der Anforderungsanalyse wurden zwei Gruppen von Attributen identifiziert, welche für die Entscheidung, ob eine künstliche Beleuchtung von Bedeutung sind: Kontextattribute und Umgebungsattribute.

Kontextattribute geben einem Ereignis einen gewissen Kontext und versuchen einen Bezug daraufhin herzustellen, aus welchem Zweck das Ereignis ausgelöst wurde. Als für die vorliegende Arbeit relevantes Kontextattribut, wurde der Zeitpunkt von Statusveränderungen definiert. Konkret die Uhrzeit am Tag, sowie der Wochentag und die Jahreszeit. Zusätzlich wurden aufgezeichnete Ereignisse und Statusveränderungen von relevanten Verbrauchern im betroffenen Raum als relevante Kontextattribute definiert. Durch den Zeitpunkt einer Veränderung ist es möglich, sich wiederholende Abläufe leichter zu erkennen und zu berücksichtigen. Durch die Aufzeichnung von bestimmten Ereignissen im betroffenen Raum ist es möglich, komplexe Aktionen der Bewohner und Bewohnerinnen in das Modell einfließen zu lassen. Das bezieht alle Ereignisse ein, die im Smart Home aufgezeichnet werden können und auf Aktionen schließen lassen, welche für eine Statusänderung in der Beleuchtung relevant sein können. Ein Beispiel hierfür wäre das Einschalten des Fernsehers, dieses Ereignis kann ein Indiz dafür sein, dass die Fernsehbeleuchtung eingeschaltet werden soll und die restliche Beleuchtung im Raum ausgeschaltet werden soll. Falls in einem Raum mehrere Lichtinstallationen vorhanden sind, so ist der Status der anderen Lichtinstallation für die Vorhersage einer Lichtinstallation ebenso eine relevante Kontextinformation. Diese Information hilft zu entscheiden, ob eine weitere Lichtquelle in der aktuellen Situation benötigt wird oder nicht. Auch der Zustand eines eventuell vorhandenen Beschattungssystems wurde als relevante Information identifiziert. Die hier relevante Information kann aber auch aus dem Helligkeitswert im Raum abgeleitet werden. Zusammengefasst sind Kontextattribute kennzeichnend für Gewohnheiten und Tagesabläufe der Bewohner und Bewohnerinnen des Smart Homes im Kontext einer Statusänderung von Beleuchtungen.

Umgebungsattribute beschreiben den Zustand der Umgebung im und rund um das Smart Home. In Anlehnung an Mahdavi, Mohammadi, Kabir und Lambeva (2008) werden darunter, für die vorliegende Arbeit, primär Attribute verstanden, welche für die Entscheidung wichtige Werte darstelle. Beispiele für solche Umgebungsattribute sind Werte wie Helligkeit, Temperatur oder Sonneneinstrahlung. Als für die Entscheidung in der vorliegenden Arbeit wichtige Umgebungsattribute wurden in der Anforderungsanalyse die Helligkeit und die Temperatur identifiziert. Je nach Temperatur und Helligkeit kann sich das Bedürfnis der Bewohner und Bewohnerinnen nach Beleuchtung unterschiedlich ausprägen. Fehlende Helligkeit ist der primäre Treiber für den Bedarf an künstlicher Beleuchtung. Auch die Unterschiede in den Tagesabläufen durch die unterschiedlichen Jahreszeiten und Übergangszeiten kann anhand dieser beiden Attribute in Kombination mit dem Zeitpunkt abgeleitet werden. Zusätzlich wurde auch die Intensität der Sonneneinstrahlung und der Winkel der Sonneneinstrahlung als relevante Umweltattribute identifiziert, die hier enthaltene Information kann teilweise aber auch über die Helligkeit abgeleitet werden.

Aus fachlicher Sicht wird die Evaluierung des Prototyps auf den Wohnbereich des für die Evaluierung verwendeten Smart Homes eingeschränkt. Konkret auf die Räumlichkeiten: Wohn- und Essbereich, Küche und Badezimmer. Andere Räumlichkeiten wie Schlafzimmer, Kinderzimmer, WC und Büro werden nicht für die Evaluierung herangezogen.

Die durch die Anforderungsanalyse ermittelten Informationen und Anforderungen an den Prototyp der vorliegenden Arbeit, bildeten die Basis des folgenden Prozessschrittes im CRISP-DM Prozess, dem Verständnis der Daten.

4.2.2 Verständnis der Daten

In diesem Prozessschritt liegt laut Abschnitt 2.2 der Fokus auf den Daten. Ziel ist es zu klären, welche Daten benötigt werden, um die fachlichen Anforderungen abzudecken und ob alle relevanten Daten in erforderlicher Qualität vorliegen.

Als erste Analyse in diesem Prozessschritt wurden die Datenquellen, die Datenstrukturen und die dazugehörenden Prozesse in der bestehenden Smart Home Instanz untersucht. Das vorliegenden Smart Home verfügt über zwei Systeme, in denen die erfassten Informationen und Ereignisse persistiert werden: Eine relationale Datenbank, in welcher Ereignisse und Statuswechsel als Schlüssel-Wert Paare je Objekt mit Zeitbezug in Tabellen gespeichert werden und eine spezielle zeitreihen Datenbank, welche fortlaufende Umgebungsattribute, wie die Temperatur, oder den Energieverbrauch als Zeitreihen persistiert. Als relationale Datenbank ist die Open Source Datenbank Software MariaDB im Einsatz, welche über SQL abgefragt werden kann. Als Zeitreihen Datenbank wird das Datenbanksystem InfluxDB eingesetzt, welches über eine eigene HTTP/REST Schnittstelle abgefragt werden kann. Beide Datenbanksysteme werden direkt von der eingesetzten Smart Home Software OpenHab mit Daten versorgt. In OpenHab gibt es dazu eigene Persistenz Regeln, mit welchen definiert werden kann, für welche Objekte im Smart Home und unter welchen Bedingungen Zustands- oder Wertveränderungen in welchem Datenbanksystem zu persistieren sind. Die im Smart Home, welches in der vorliegenden Arbeit eingesetzt wird, vorhandene Konfiguration sieht vor, dass jede Veränderung eines Zustands, eines Objektes oder eine Wertveränderung eines Objektes persistiert wird. Für fortlaufende numerische Sensorwerte ist das Zeitreihen System InfluxDB das zuständige Datenbanksystem. Für die Überwachung von Zuständen von Objekten wie Beleuchtungen oder andere Verbraucher im Smart Home, ist das relationale System MariaDB das zuständige Datenbanksystem. Hauptgrund für diese Trennung der Zuständigkeiten ist, dass die für Zeitreihen optimierte Datenbank InfluxDB zwar sehr gut und effizient mit numerischen Werten umgehen kann, für klassifizierende Attribute aber einige Nachteile mit sich bringt. Veränderungen an Werten oder Zuständen eines Objektes werden direkt bei jeder Änderung persistiert und historisiert. Somit ist während des Aufzeichnungszeitraumes eine zeitlich lückenlose Nachvollziehbarkeit der Werte oder Zustände eines Objektes gegeben.

Anschließend wurden die Räumlichkeiten für die Evaluierung laut Anforderungsanalyse untersucht, um zu erheben, welche relevanten Kontext- und Umgebungsattribute in diesen Räumlichkeiten vorhanden und in den Datenquellen persistiert sind. Dabei wurden wie in Tabelle 4 ersichtlich, auch einige globale Attribute identifiziert, welche für die Vorhersage in jedem der einbezogenen Räume, relevant sind. Neben der Außentemperatur, der Sonneneinstrahlung und der Helligkeit, wurden weitere Attribute identifiziert, welche einer nähren Beschreibung bedürfen. Der aktuelle Stromverbrauch kann als Kontextattribut gesehen

werden, um in Kombination mit dem Zeitpunkt, eine Verbindung zu sich wiederholenden Abläufen und komplexen Aktionen im Smart Home herzustellen. Die Präsenz der Bewohner und Bewohnerinnen lässt ebenso Schlüsse auf Tagesabläufe zu und ist daher auch als Kontextattribute zu sehen. Die Präsenz der Bewohner und Bewohnerinnen im Smart Home wird mittels deren Smartphones und deren Konnektivität zum WLAN im Smart Home aufgezeichnet und historisiert. Im Smart Home existiert zusätzlich ein virtuelles Nachtmodus Statusobjekt welches Funktionen, wie die Außenbeleuchtung oder Bewegungsmelder steuert. Dieser Nachtmodus wird aus einer Kombination aus Helligkeitswerten, sowie dem, am jeweiligen Tag, errechneten Sonnenzyklus, am Standort des Smart Homes, gesteuert und hat Einfluss auf andere Funktionen im Smart Home.

Globale Attribute		
Attribute	Persistenz	Datentyp
Außentemperatur	InfluxDB	Numerischer Messwert in Grad Celsius
Aktueller Stromverbrauch L1	InfluxDB	Numerische Messwerte in Watt
Aktueller Stromverbrauch L2	InfluxDB	
Aktueller Stromverbrauch L3	InfluxDB	
Sonneneinstrahlung	InfluxDB	Numerischer Messwerte in Watt
Helligkeit	InfluxDB	Numerischer Messwert in Lux
Präsenz	MariaDB	Boolescher Wert je Bewohner und Bewohnerin
Smart Home Nachtmodus	MariaDB	Boolescher Wert

Tabelle 4: Globale Attribute im Smart Home

Lokale Attribute: Küche		
Attribute	Persistenz	Datentyp
Beleuchtung Küche	MariaDB	Boolescher Wert
Unterschrankbeleuchtung	MariaDB	Boolescher Wert
Raumtemperatur	InfluxDB	Numerischer Messwert in Grad Celsius
Beleuchtung Esstisch	MariaDB	Boolescher Wert
Beleuchtung Fernseher (eingefügt laut 4.2.4)	MariaDB	Boolescher Wert
Beleuchtung Spielecke (eingefügt laut 4.2.4)	MariaDB	Boolescher Wert

Tabelle 5: Lokale Attribute: Küche

Für den Raum Küche wurden lokale Kontext- und Umgebungsattribute, wie in Tabelle 5 ersichtlich, identifiziert. Hier wurden die zwei Beleuchtungsobjekte selbst und die Raumtemperatur identifiziert. Für den Helligkeitswert und andere Attribute muss auf globale Attribute zurückgegriffen werden. Speziell der aktuelle Stromverbrauch kann hier eine Rolle spielen, da in der Küche sehr viele Verbraucher angesiedelt sind. Zusätzlich wurde die Esstisch Beleuchtung als Attribut identifiziert. Der Esstisch befindet sich zwar nicht direkt in der Küche, der Status der Beleuchtung über diesem kann, durch die offene Bauweise des Wohnraumes,

einen Einfluss auf den Beleuchtungsbedarf in der Küche haben. Durch eine, in Abschnitt 4.2.4 beschriebene, Iteration im CRISP-DM Prozess, wurde eine Analyse der für die Küche relevanten Attribute durchgeführt. Dabei wurde festgestellt, dass der Status der Beleuchtungen Fernseher und Spielecke einen großen Informationsgewinn zur Vorhersage des Status der Beleuchtung in der Küche darstellen.

Für den Raum Wohn- und Esszimmer wurden lokale Kontext- und Umgebungsattribute, wie in Tabelle 6 ersichtlich, identifiziert. In diesem zentralen Wohnbereich gibt es vier Beleuchtungsquellen und einen Helligkeitssensor im Raum. Als relevante Statusinformation wurde hier zusätzlich der Status des Fernsehers identifiziert.

Lokale Attribute: Wohn- und Esszimmer		
Attribute	Persistenz	Datentyp
Beleuchtung Fernseher	MariaDB	Boolescher Wert
Beleuchtung Spielecke	MariaDB	Boolescher Wert
Beleuchtung Mitte	MariaDB	Boolescher Wert
Beleuchtung Esstisch	MariaDB	Boolescher Wert
Raumtemperatur	InfluxDB	Numerischer Messwert in Grad Celsius
Status Fernseher	MariaDB	Boolescher Wert
Helligkeit Wohnzimmer	InfluxDB	Numerischer Messwert in Lux

Tabelle 6: Lokale Attribute Wohn- und Esszimmer

Für den Raum Badezimmer wurden lokale Kontext- und Umgebungsattribute, wie in Tabelle 7 ersichtlich, identifiziert. Hier wurden die zwei Beleuchtungsobjekte selbst und die Raumtemperatur identifiziert. Für den Helligkeitswert und andere Attribute muss auf globale Attribute zurückgegriffen werden. Durch eine, in Abschnitt 4.2.4 beschriebene, Iteration im CRISP-DM Prozess wurde eine Analyse der für das Bad relevanten Attribute durchgeführt. Dabei wurde festgestellt, dass der Status der Beleuchtungen der Küche, des Esstisches und die Temperatur im Wohnraum einen großen Informationsgewinn zur Vorhersage des Status der Beleuchtung im Bad darstellen.

Lokale Attribute: Badezimmer		
Attribute	Persistenz	Datentyp
Beleuchtung Badezimmer	MariaDB	Boolescher Wert
Beleuchtung Spiegel	MariaDB	Boolescher Wert
Raumtemperatur Bad	InfluxDB	Numerischer Messwert in Grad Celsius
Beleuchtung Esstisch (eingefügt laut 4.2.4)	MariaDB	Boolescher Wert
Beleuchtung Küche (eingefügt laut 4.2.4)	MariaDB	Boolescher Wert
Raumtemperatur Wohnraum (eingefügt laut 4.2.4)	InfluxDB	Numerischer Messwert in Grad Celsius

Tabelle 7: Lokale Attribute Badezimmer

Alle betrachteten Räume verfügen über eine entsprechende Beschattung inklusive Smart Home Anbindung. Die Datenqualität dieser Anbindung ist aber nicht ausreichend, um den Zustand der Beschattung mitberücksichtigen zu können. Es wurde auch festgestellt, dass nicht jeder der Räume über einen eigenen Helligkeitssensor verfügt. Diese Information kann aber mit Hilfe eines globalen Helligkeitssensors und der Uhrzeit, in Kombination mit dem Sonnenzyklus zumindest näherungsweise errechnet werden. Der Zustand der Beschattung kann in dieser Berechnung nicht berücksichtigt werden.

Der Zeitbezug eines Attributwertes wurde als sehr wichtiges Attribut für den vorliegenden Anwendungsfall identifiziert. Durch dieses Attribut ist es möglich, den Bezug zwischen allen anderen Attributwerten herzustellen. Außerdem ist es durch dieses Attribut möglich, auf Tagesabläufe und wiederkehrende Vorgänge schließen zu können. Für das Data Mining Modell der vorliegenden Arbeit wurden die, in Tabelle 8 ersichtlichen, Attribute für den Zeitbezug identifiziert. Die Uhrzeit des Tages stellt das Kernattribut dar. Diese Uhrzeit wird in vergangene Minuten des Tages abgebildet. Dieses Attribut wird anschließend noch in zwei Verdichtungsstufen abgebildet und in das Data Mining Modell eingebracht. Einmal als Wochentag und einmal als Tageszeitattribut, welches den Tag in die fachlichen Zeitabschnitte Nacht, Morgen, Mittag, Nachmittag und Abend einteilt. Aus fachlicher Sicht würde auch eine weitere Verdichtung in Monate oder Jahreszeiten Sinn ergeben. Der Evaluierungszeitrum in der vorliegenden Arbeit ist aber nicht lange genug, damit diese Attribute Auswirkungen zeigen würden.

Globale Attribute: Zeitbezug	
Attribute	
Minute des Tages	Die Uhrzeit abgebildet als die Anzahl vergangener Minuten am jeweiligen Tag (von 0 bis 1440)
Tageszeit	Attribut mit folgenden Klassen, klassifiziert nach der Uhrzeit: • 22-24 und 0-5 Uhr: Nacht • 5-10 Uhr: Morgen • 10-14 Uhr: Mittag • 14-19 Uhr: Nachmittag • 19-22 Uhr. Abend
Wochentag	Der Wochentag abgebildet als Attribut mit einer Klasse je Wochentag (Montag – Sonntag)

Tabelle 8: Attribute für den Zeitbezug

Zusammengefasst wurden 30 lokale und globale Attribute identifiziert, welche für das Training des Data Mining Modells in der vorliegenden Arbeit eingebunden werden. Für jede Beleuchtung, deren Status vorhergesagt werden soll, muss ein eigenes Data Mining Modell trainiert werden, dabei wird die Auswahl der relevanten Attribute des Trainingsdatensatzes, eines jeden Data Mining Modells, aus diesen 30 Attributen zusammengestellt. Als wichtigster Aspekt für diesen Anwendungsfall wurde der zeitliche Bezug festgestellt. Eine jede Instanz im

Trainingsdatensatz muss einen zeitlichen Bezug für jedes beinhaltete Attribut herstellen. Der zeitliche Bezug stellt das zentrale Element einer jeden Instanz im Datensatz dar.

4.2.3 Datenaufbereitung

In diesem Prozessschritt liegt, laut Abschnitt 2.2, der Fokus auf der Extraktion und Aufbereitung der Daten, damit sie für das Erstellen des Data Mining Modells leicht handhabbar sind. Wie in Abschnitt 4.2.2 beschrieben, ist es für den Anwendungsfall der vorliegenden Arbeit notwendig, mehr als ein Data Mining Modell zu erstellen. Diesen Data Mining Modellen liegt derselbe Datensatz für das Training zugrunde. Je nach Data Mining Modell werden unterschiedliche Attribute aus diesem Datensatz ausgewählt. Somit ist es ausreichend einen großen Datensatz für alle Data Mining Modelle zu erstellen und nicht für jedes Data Mining Modell einen eigenen Datensatz zu erstellen.

In den beiden Datenquellen MariaDB und InfluxDB sind alle relevanten Informationen in Form von Ereignissen mit Zeitbezug persistiert. Diese Form der Daten ist für das Training eines Data Mining Modells nicht gut geeignet, weil keine vollständigen Instanzen erstellt werden können, wenn nicht für alle relevanten Attribute zum gleichen Zeitpunkt ein Ereignis persistiert wurde. In Abschnitt 4.2.2 wurde der Zeitbezug als eines der wichtigsten Attribute identifiziert. Der Zeitbezug stellt den Kernzusammenhang zwischen allen Attributwerten einer Instanz dar. Es wurde auch festgestellt, dass durch die ereignisbezogene Persistenz der Werte eine lückenlose Historie der Werte eines Attributes, im Aufzeichnungszeitraum, möglich ist. Diese zwei Erkenntnisse werden die Basis für die Extraktion und das Format des Datensatzes bilden. Die Daten müssen in ein Format extrahiert und transformiert werden, in dem die Attributwerte einer jeden Instanz, den Zustand im Smart Home zu einem bestimmten Zeitpunkt wiederspiegeln.

Die folgend beschriebene Vorgehensweise wurde für die Extraktion der Daten in der vorliegenden Arbeit festgelegt. Für die Extraktion der Daten muss ein Zeitraum und eine zeitliche Auflösung festgelegt werden. Anschließend wird in dieser zeitlichen Auflösung der komplette Zeitraum durchlaufen und für jede Iteration eine Instanz im Datensatz angelegt. Diese Instanzen werden anschließend mit den entsprechenden Attributwerten, welche zum Zeitpunkt der jeweiligen Instanz im Smart Home gültig waren, befüllt. Falls unvollständige Daten vorliegen, wird der zuletzt bekannte Attributwert übernommen. Durch diese Vorgehensweise ist bei kurzen Ausfällen einzelner Sensoren eine vollständige Instanz gewährleistet. Als zeitliche Auflösung wurde eine Minute definiert, diese Auflösung ermöglicht eine sehr genaue Abbildung der Zustände im Smart Home. Durch diese Vorgehensweise wird die Extraktion und die Transformation der Daten in einem Schritt abgebildet und ein, für das Training des Data Mining Modells fertiger Datensatz, ist vorhanden.

Der Datensatz wird in Form von kommaseparierten Dateien exportiert und am Dateisystem des Servers abgelegt, an dem das Data Mining Modell erstellt wird. Dabei wird für jeden Tag eine eigene Datei angelegt. Diese Vorgehensweise hat den Vorteil, dass immer nur die Zeitraumdifferenz seit der letzten Extraktion exportiert werden muss, um einen vollständigen Datensatz über den gesamten Zeitraum zu erhalten.

Für die Anforderungen der vorliegenden Arbeit ist es notwendig, dass diese Extraktion der Daten automatisiert, mit erweitertem Zeitraum wiederholt werden kann. Für ein neuerliches Training der Data Mining Modelle ist es erforderlich, einen aktualisierten Trainingsdatensatz aufzubauen.

4.2.4 Modellierung und Evaluierung des Data Mining Modells

In diesem Prozessschritt liegt, laut Abschnitt 2.2, der Fokus auf der Erstellung des Data Mining Modells. Die Auswahl des zu verwendenden Data Mining Verfahrens wird auch als Teil dieses Prozessschrittes gesehen werden. Da in der vorliegenden Arbeit das zu verwendende Data Mining Verfahren bereits in Abschnitt 3.10 auf das Data Mining Verfahren Entscheidungsbäume festgelegt wurde, wird die Auswahl hier nicht weiter berücksichtigt. Für den Anwendungsfall der vorliegenden Arbeit und der in Abschnitt 4.2.1 und 4.2.2 getroffenen Auflistung von Räumen und Beleuchtungen, welche im Prototyp berücksichtigt werden sollen, ist es notwendig, in Summe sechs Data Mining Modelle zu erstellen. Es wird je ein Data Mining Modell für die Vorhersage der folgenden Beleuchtungen erstellt:

- Modell 1: Badezimmer – Hauptlicht
- Modell 2: Küche - Hauptlicht
- Modell 3: Wohnbereich – Beleuchtung Esstisch
- Modell 4: Wohnbereich – Beleuchtung Fernseher
- Modell 5: Wohnbereich – Beleuchtung Mitte
- Modell 6: Wohnbereich – Beleuchtung Spieleecke

Die sechs Data Mining Modelle werden aus demselben Datensatz trainiert. Je nach Data Mining Modell werden die benötigten Attribute aber spezifisch ausgewählt und berücksichtigt. Die Basis für diese Auswahl bildet die Aufstellung der lokalen Attribute je Raum, sowie die globalen Attribute in Abschnitt 4.2.2.

Für die Umsetzung des Prototypen wurde der, laut Wu et. al (2007), weit verbreitete Algorithmus, C4.5 aus der Familie der Entscheidungsbäume ausgewählt. Dieser Algorithmus zählt zu den, am meisten genutzten Entscheidungsbaum Algorithmen und wird von Weka in der Open Source Implementierung J48 unterstützt.

Für das Finden von geeigneten Attributwerten für das Training des Data Mining Modells mit dem gewählten Algorithmus und die Evaluierung der Attributauswahl, wurde die Weka Software verwendet. Der Datensatz wurde in Weka geladen und es wurden Modelle mit dem Datensatz und dem gewählten Algorithmus gebaut. Ziel dabei war es, möglichst gute Vorhersagewerte zu erreichen und dabei eine Überanpassung des Data Mining Modells an die Testdaten zu vermeiden. Gute Vorhersagewerte können anhand einer hohen Präzision und Trefferquote erkannt werden. Überanpassung kann, durch Training und Evaluierung mit zwei unterschiedlichen Datensätzen, erkannt werden. Um die Werte zu erheben, wurden für jedes Data Mining Modell zwei Validierungen durchgeführt. Zuerst eine Kreuzvalidierung mit zehn

Partitionen um die Präzision und die Trefferquote zu ermitteln, anschließend wird mit einem eigenen Datensatz, welcher nicht zum Training verwendet wurde, eine Validierung mit allen enthaltenen Instanzen durchgeführt. Die beiden Datensätze müssen dabei aus derselben Population kommen. Der Unterschied in den Vorhersagewerten zwischen den zwei Durchläufen lässt einen Rückschluss darauf zu, ob eine Überanpassung vorliegt oder nicht. Es wurden zwei Datensätze für diese Evaluierung aus dem Smart Home erstellt. Der Zeitraum für das Training umfasst den Zeitraum von Mitte Juli bis Ende August 2017. Somit steht bei einer Aufteilung ein gleich großer Trainings- und Evaluierungsdatensatz, welcher je einen Zeitraum von drei Wochen abdeckt, für Training und Validierung zur Verfügung. Die Analyse der ersten Data Mining Modelle zeigte, dass die gewichtete Gesamtvorhersagequalität sehr gut war. Bei genauerer Betrachtung wurde aber ersichtlich, dass die ermittelte Vorhersagequalität der Klasse „Beleuchtung Aus“ (Wert „OFF) sehr gut abgeschnitten hat, die Vorhersage der Klasse „Beleuchtung Ein“ (Wert „ON“) jedoch nicht gut abgeschnitten hat. Durch eine Analyse der Daten ließ sich dieses Verhalten erklären. Im Datensatz ist ein großes Ungleichgewicht in der Klassenverteilung zu erkennen. So wurde im Trainingsdatensatz zum Beispiel für das Attribut „Beleuchtung Esstisch“, ein Verhältnis von 30 Instanzen mit der Klasse „Beleuchtung Aus“ zu einer Instanz mit der Klasse „Beleuchtung Ein“ gefunden. Dieses Ungleichgewicht ergibt sich aus dem vorliegenden Anwendungsfall. Am Großteil der Zeit an einem Tag, ist eine Beleuchtung nicht eingeschalten. Die unterschiedlichen Vorhersagequalitäten je Klasse sind für den vorliegenden Anwendungsfall als wichtige Information eingestuft worden. Aus diesem Grund wurde entschieden, in der Auswertung die Vorhersagequalität, die Präzision und die Trefferquote getrennt für beide Klassen auszuwerten. Diese Art der Auswertung entspricht der Auswertung der Konfusionsmatrix.

Modell	TP „OFF“	FP „OFF“	Präzision „OFF“	Trefferquote „OFF“	TP „ON“	FP „ON“	Präzision „ON“	Trefferquote „ON“
1	99,4%	51,4%	98,7%	99,4%	48,6%	0,6%	68,1%	56,7%
2	98,8%	21,5%	97,8%	98,8%	78,5%	1,2%	87,1%	78,5%
3	99,4%	19,6%	99%	99,4%	80,4%	0,6%	86,9%	80,4%
4	99,6%	12,5%	99,2%	99,6%	87,5%	0,4%	93,2%	87,5%
5	99,6%	5,5%	99,5%	99,5%	94,5%	0,4%	95,6%	94,5%
6	99,9%	9,4%	99,7%	99,9%	90,6%	0,1%	96,2%	90,6%

Tabelle 9: Trainingsergebnisse der Data Mining Modelle mit Kreuzvalidierung

Der erste Trainingsdurchlauf wurde mit den von Weka vorgesehenen Standardparametern für den C4.5 Algorithmus durchgeführt. Die Ergebnisse des Durchlaufes mit einer zehnfachen Kreuzvalidierung, sind in Tabelle 9 dargestellt. Die Ergebnisse des Durchlaufes mit einem 50% Split des Datensatzes, sind in Tabelle 10 abgebildet. Auffällig in beiden Durchläufen ist, dass die Vorhersagequalität für Klasse „Beleuchtung Aus“ (Wert „OFF“ in der Tabelle) durchgehend sehr hoch ist und es auch zwischen den zwei Durchläufen nur geringe Unterschiede gibt.

Wohingegen bei der Vorhersagequalität für die Klasse „Beleuchtung Ein“ (Wert „ON“ in der Tabelle) sowohl innerhalb der Durchläufe, als auch zwischen den beiden Durchläufen größere Unterschiede identifiziert wurden. Vor allem die Unterschiede zwischen den Durchläufen der Modelle 1-3 deuten auf keine guten Vorhersagewerte des Modells hin, da die Vorhersagequalität beim zweiten Durchlauf stark gesunken ist.

Modell	TP „OFF“	FP „OFF“	Präzision „OFF“	Trefferquote „OFF“	TP „ON“	FP „ON“	Präzision „ON“	Trefferquote „ON“
1	99,8%	79,5%	98,0%	99,8%	20,5%	73,2%	20,5%	32,0%
2	98,4%	32,4%	96,7%	98,4%	67,6%	1,6%	81,1%	67,7%
3	99,3%	32,6%	98,4%	99,3%	67,4%	0,7%	82,6%	74,2%
4	99,1%	13,3%	99,2%	99,1%	86,7%	0,9%	85,5%	86,7%
5	98,9%	10,1%	99,0%	98,9%	89,9%	1,1%	88,7%	89,9%
6	99,7%	8,8%	99,7%	99,7%	91,2%	0,3%	90,8%	91,2%

Tabelle 10: Trainingsergebnisse der Data Mining Modelle mit getrenntem Validierungsdatensatz

Auf Basis der Ergebnisse dieser beiden Durchläufe, wurden Ansätze zur Verbesserung definiert, um die Vorhersagequalität zu erhöhen und die Überanpassung zu reduzieren. Diese Ansätze wurden für die Modelle 1, 2 und 3 erprobt und evaluiert, da hier der größte Verbesserungsbedarf bestand.

Der erste Verbesserungsansatz bestand darin, die ausgewählten Attribute zu hinterfragen und die Attributauswahl zu prüfen. Im Sinne des CRISP-DM Prozesses wäre dieser Ansatz, wie in Abschnitt 2.2 beschrieben, ein Schritt zurück zum Prozessschritt Datenaufbereitung. Konkret wurden alle in Abschnitt 4.2.3 definierten Attribute auf ihre Relevanz für die Vorhersage für das jeweilige Modell analysiert. Das Vorgehen für die Analyse der Attribute und der Auswahl der relevantesten Attribute wurde von Tirelli und Pessani (2011) übernommen. Tirelli und Pessani haben verschiedene in Weka verfügbare Methoden zur Auswahl von Attributen für die C4.5 Implementierung evaluiert. Für die vorliegende Arbeit wurde dafür konkret die Bewertungsmethode „InfoGainAttributeEval“, in Kombination mit der Suchmethode „Ranker“ angewandt, welche auch eine der Varianten ist, welche von Tirelli und Pessani (2011) evaluiert wurden. Bei der Bewertungsmethode „InfoGainAttributeEval“ werden Attribute anhand des Informationsgewinns, den sie im Datensatz für die jeweilige Klassifizierung beitragen, bewertet. Diese Bewertungsmethode wird von der Suchmethode „Ranker“ verwendet, um für die Reihung aller Attribute die entsprechende Bewertung zu berechnen (Witten, Frank, Hall, & Pal, 2016). Für das Data Mining Modell 1 wurde mit dieser Analyse erkannt, dass die drei Attribute „Beleuchtung Küche“, „Beleuchtung Esstisch“ und „Temperatur Wohnraum“ einen Informationsgewinn für dieses Data Mining Modell darstellen. Als Folge dieser Analyse wurden die drei Attribute zum Data Mining Modell 3, als Attribut hinzugefügt. Für das Data Mining Modell 2 wurde mit dieser Analyse erkannt, dass die Attribute „Beleuchtung Fernseher“ und „Beleuchtung Spielecke“, welche beide für dieses Modell nicht angedacht waren, einen großen Informationsgewinn bringen. Als Folge dieser Analyse wurden die beiden Attribute zum Data

Mining Modell 2, als Attribut hinzugefügt. Für das Data Mining Modell 3 wurde mit dieser Analyse festgestellt, dass das Attribut „Beleuchtung Küche“, welches für dieses Modell nicht angedacht war, einen Informationsgewinn bringt. Als Folge dieser Analyse wurde das Attribut zum Data Mining Modell 3 als Attribut hinzugefügt.

Als zweiter Verbesserungsansatz wurden die Parameter des Algorithmus J48 nachjustiert, um bessere Vorhersagewerte zu erhalten. Dieser Ansatz wurde durch Ausprobieren verschiedener Parameterkombinationen umgesetzt. Konkret wurden die Parameter, welche eine Überanpassung des Modells an die Daten reduzieren oder verhindern sollten, dahingehend optimiert, um bei geringer Überanpassung sehr hohe Vorhersagewerte zu bekommen. Die Parametereinstellung, welche durch diesen Verbesserungsansatz erstellt wurde, ist in Tabelle 11 ersichtlich. Bei der Umsetzung ist speziell aufgefallen, dass die Mechanismen „subtreeRaising“ und „reducedErrorPruning“, welche beide einer Überanpassung des Data Mining Modells entgegenwirken sollen, in diesem Anwendungsfall nicht sinnvoll waren, da sich keinerlei Unterschied in den Vorhersagewerten ergeben haben.

Parameter	Wert	Beschreibung
binarySplits	False	Steuert, ob der Baum als binärer Baum aufgebaut wird oder nicht. (Weka Java Doc, 2017).
collapseTree	False	Steuert, ob der Algorithmus, Teilbäume automatisch entfernt wenn bessere Teilbäume gefunden wurden (Weka Java Doc, 2017).
confidenceFactor	0,2	Definiert den Konfidenzwert der erreicht werden muss, damit der Algorithmus Maßnahmen gegen Überanpassung einleitet (Weka Java Doc, 2017).
minNumObj	10	Definiert die minimale Anzahl von Instanzen, die notwendig sind, damit ein eigener Knoten im Baum für die Entscheidung angelegt wird (Weka Java Doc, 2017).
numDecimalPlaces	0	Definiert die Anzahl der Nachkommastellen, welche bei numerischen Werten verwendet werden (Weka Java Doc, 2017).
numFolds	10	Steuert die Anzahl der Daten, die für Maßnahmen gegen Überanpassung verwendet werden (Weka Java Doc, 2017).
reducedErrorPruning	False	Auswahl des Algorithmus gegen Überanpassung (Weka Java Doc, 2017).
subtreeRaising	False	Steuert, ob Teilbäume im Zuge einer Maßnahme gegen die Überanpassung in der Hierarchie nach oben geschoben werden dürfen (Weka Java Doc, 2017).
useLaplace	False	Steuert, ob die Kanten des Baumes mit La Place Berechnungen abgerundet werden sollen (Weka Java Doc, 2017).
useMDLcorrection	True	Auswahl des Algorithmus für die Ermittlung eines Splits für numerische Parameter (Weka Java Doc, 2017).

Tabelle 11: Parametereinstellungen J48 Algorithmus

Nach der Umsetzung dieser beiden Verbesserungen wurden die Modelle erneut mit angepassten Parametern trainiert. Wie in den Ergebnissen des Durchlaufes mit Kreuzvalidierung in Tabelle 12 ersichtlich ist, konnten die Vorhersagewerte für die Klasse „Beleuchtung ein“ der Modelle 1-3 verbessert werden. Aber auch die bereits hohen Vorhersagewerte für die Modelle 4-6, konnten noch weiter verbessert werden. Wie in Tabelle 13 ersichtlich, konnte auch beim zweiten Durchlauf, mit einem getrennten Validierungsdatensatz, eine Verbesserung bei den Modellen 1-3 identifiziert werden. Auch hier war es möglich, die Vorhersagewerte der Modelle 4-6 noch etwas zu verbessern.

Modell	TP „OFF“	FP „OFF“	Präzision „OFF“	Trefferquote „OFF“	TP „ON“	FP „ON“	Präzision „ON“	Trefferquote „ON“
1	99,6%	37,9%	99%	99,6%	62,1%	0,4%	78,4%	62,1%
2	99,0%	9,8%	99,0%	99,0%	90,2%	1,0%	90,8%	90,2%
3	99,5%	14,9%	90,5%	99,5%	85,1%	0,5%	85,1%	87,7%
4	99,6%	8,6%	99,5%	99,6%	91,4%	0,4%	94,0%	91,4%
5	99,7%	4,0%	99,6%	99,7%	96,0%	0,3%	96,7%	96,0%
6	99,9%	3,5%	99,9%	99,9%	96,5%	0,1%	97,5%	96,5%

Tabelle 12: Zweiter Durchlauf - Trainingsergebnisse der Data Mining Modelle mit Kreuzvalidierung

Modell	TP „OFF“	FP „OFF“	Präzision „OFF“	Trefferquote „OFF“	TP „ON“	FP „ON“	Präzision „ON“	Trefferquote „ON“
1	99,6%	54,4%	98,6%	99,6%	45,6%	0,4%	73,8%	45,6%
2	98,8%	15,5%	98,4%	98,8%	84,5%	1,2%	87,6%	84,5%
3	99,1%	17,0%	99,1%	99,1%	83,0%	0,9%	82,5%	83%
4	99,6%	9,6%	99,4%	99,6%	90,4%	0,4%	92,8%	90,4%
5	99,5%	7,2%	99,3%	99,5%	92,8%	0,5%	94,7%	92,8%
6	99,9%	10,3%	99,7%	99,9%	89,7%	0,1%	97,3%	89,7%

Tabelle 13: Zweiter Durchlauf - Trainingsergebnisse der Data Mining Modelle mit getrenntem Validierungsdatensatz

Alle Vorhersagewerte für die Klasse „Beleuchtung Aus“ liegen mit ermittelten Einstellungen, im schlechteren der beiden Versuche, bei über 98% korrekten Vorhersagen. Die Vorhersagewerte für die Klasse „Beleuchtung Ein“ liegen bei allen Data Mining Modellen, außer dem Modell 1, bei über 83% korrekter Vorhersagen. Das Data Mining Modell 1 für die Vorhersage des Zustandes der Beleuchtung im Badezimmer, konnte mit den ausgewählten Daten nicht auf über 45,6% korrekter Vorhersagen gesteigert werden. Für dieses Data Mining Modell wären weitere Attribute notwendig, um die Vorhersagen zu verbessern. Diese Attribute sind jedoch im Smart Home aktuell nicht vorhanden. Die hier ermittelte Konfiguration der Attributen in den Datensätzen der einzelnen Data Mining Modelle, sowie die ermittelte Parametereinstellung des

J48 Algorithmus, wurden in weiterer Folge für die Integration der Data Mining Modelle in den Prototypen verwendet.

4.3 Implementierung des Prototyps

In diesem Abschnitt wird der Aufbau, die Struktur und die Funktionsweise des erstellten Prototyps, sowie die Integration in die bestehende Smart Home Infrastruktur beschrieben. Aus Sicht des, in Abschnitt 2.2 beschriebenen, CRISP-DM Prozesses, handelt es sich um den Prozessschritt Verwendung des Data Mining Modells. Zuerst wird ein grober Überblick über die Lösungsarchitektur gegeben und die einzelnen Komponenten beschrieben. Anschließend wird auf die zwei Kernprozesse des Prototyps, das Training eines neuen Data Mining Modells und die Nutzung eines trainierten Data Mining Modells für Vorhersagen, eingegangen und der Ablauf dieser beschrieben.

4.3.1 Lösungsarchitektur

Der erstellte Prototyp wurde in das bestehende Smart Home des Autors der vorliegenden Arbeit integriert. Die Softwarekomponenten dieses Smart Homes sind in einer virtualisierten Umgebung auf einem kleinen Server im Smart Home installiert. Dabei wird für den Betrieb der eigentlichen Smart Home Software OpenHab, sowie dem für die Interaktion zwischen den Komponenten im Smart Home zuständigen MQTT Server, je ein eigener Docker Container für die Virtualisierung eingesetzt. Die verwendeten Datenbanksysteme MariaDB und InfluxDB sind in jeweils eigenen virtuellen Maschinen aufgesetzt und in Betrieb. Für den Prototyp der vorliegenden Arbeit werden zwei neue Komponenten erstellt und in diese Architektur wie in Abbildung 3 beschrieben integriert.

Dabei wurde eine Data Mining Komponente, welche wie in Abschnitt 4.1 entschieden, auf Spring Boot basierend, entwickelt und mittels eines Docker Containers in die bestehende Infrastruktur integriert. Die Kommunikation dieser Komponente mit den anderen Komponenten des Prototyps und des Smart Home wird über den bestehenden MQTT Bus umgesetzt. Die Data Mining Komponente ist zuständig für das Erstellen und die Verwaltung der, im Prototyp verwendeten Data Mining Modelle, sowie für das Abarbeiten von eingehenden Vorhersageanfragen von der Client Komponente und dem damit verbundenen Auslösen von Statusänderungen von Beleuchtungen über die Smart Home Applikation OpenHab. Die Data Mining Komponente ist außerdem für die für die Datenextraktion aus den, im Smart Home eingesetzten und von OpenHab mit Daten versorgten, Datenbanken und dem damit verbundenen Erzeugen von Trainingsdatensätzen zuständig.

Für die Präsenzerkennung und den damit verbundenen Anstoßen von neuen Vorhersageereignissen in der Data Mining Komponente, wurde eine neue Client Komponente in Form eines Android Apps entwickelt, welche auf den Smartphones der Bewohner und Bewohnerinnen betrieben wird. Diese App kommuniziert mit der Data Mining Komponente bei einem neuen Vorhersageereignis über den MQTT Bus und stößt so neue Vorhersagen bei der

Data Mining Komponente an. Die Hauptaufgabe der Client Komponente besteht darin, permanent nach Bluetooth Geräten in Reichweite zu suchen und zu prüfen, ob einer der eingesetzten Bluetooth Beacons in Reichweite kommt oder diese verlässt. In den Räumen im Smart Home, welche Bestandteil der Fallstudie sind, wurden Bluetooth Beacons installiert, welche ständig kleine Bluetooth Signale mit geringer Reichweite aussenden. Die Reichweite der Signale wurde je Bluetooth Beacon so konfiguriert, dass sie genau den jeweiligen Raum abdecken und darüber hinaus nicht ausreichend stark sind, um eine Verbindung mit einem Smartphone aufbauen zu können. Durch das Betreten oder Verlassen eines Raumes, durch einen Bewohner oder eine Bewohnerin mit entsprechendem Smartphone, wird durch das Client App die Präsenz eines Bluetooth Beacon erkannt oder diese verloren. Auf Basis dieser beiden Ereignisse wird von der Client Applikation, in Kombination mit dem eindeutigen Namen des Bluetooth Beacons, ein Vorhersageereignis für die Beleuchtungsquellen, im entsprechenden Raum, über den MQTT Bus in der Data Mining Komponente ausgelöst.

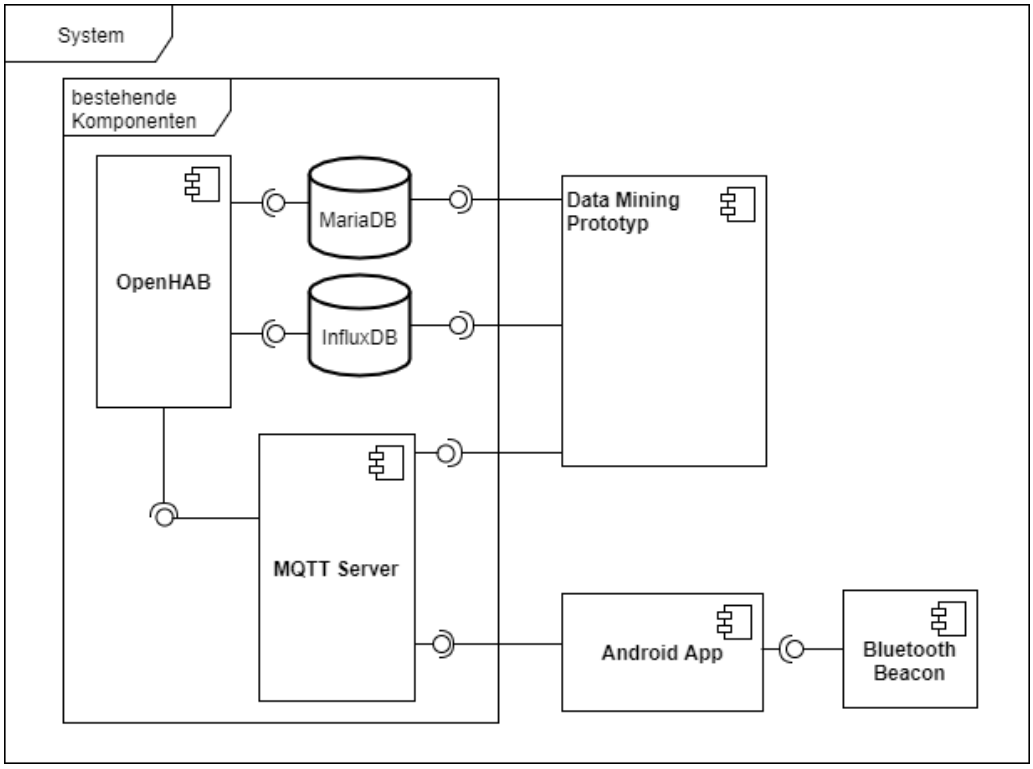

Abbildung 3: Lösungsarchitektur des Prototyps

Mit Ausnahme der Datenextraktion findet jede Interaktion zwischen den Komponenten des Prototyps und der Komponenten des Smart Homes über den MQTT Bus statt. Für die Datenextraktion aus den Datenbanken, wurde aus Geschwindigkeitsgründen, auf den direkten Zugriff auf die Datenbanksysteme mit der jeweilig angebotenen Schnittstelle zurückgegriffen.

In den folgenden zwei Abschnitten werden die zwei Anwendungsfälle des Prototyps, das Training von Data Mining Modellen, sowie das Treffen von Vorhersagen detaillierter beschrieben.

4.3.2 Trainingsprozess

Der Anwendungsfall des Trainingsprozesses umfasst die Erstellung und die Verwaltung von Data Mining Modellen in der Data Mining Komponente. Wie in Abbildung 4 ersichtlich, werden zu Beginn die Daten, die in den übergebenen Zeitraum fallen und noch nicht im bereits exportierten Datenbestand sind, aus den Datenbanken des Smart Home exportiert und zu den vorhandenen Trainingsdaten hinzugefügt. Anschließend werden die folgenden Prozessschritte sequentiell für jedes der sechs zu erstellenden Data Mining Modelle durchlaufen. Zuerst wird aus dem gesamten Datenbestand, der in den übergebenen Zeitraum fällt, ein Weka Datensatz erzeugt welcher genau die Attribute enthält, die für das jeweilige Data Mining Modell notwendig sind und der J48 Algorithmus mit den in Abschnitt 4.2.4 ermittelten Parametern initialisiert. Anschließend wird das Data Mining Modell mit dem erstellten Datensatz und dem initialisierten Algorithmus trainiert und mit einer zehnfachen Kreuzvalidierung validiert. Sowohl das erstellte Datenmodell, als auch das Evaluierungsergebnis werden zum Abschluss persistiert.

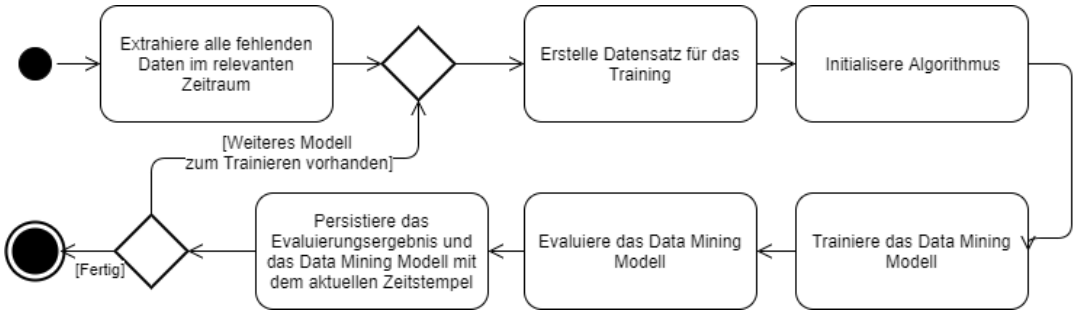

Abbildung 4: Aktivitätsdiagramm Trainingsprozess

Das erstellte Data Mining Modell wird mit dem Erstellungszeitpunkt versioniert. Auf diese Weise ist es für den Vorhersageprozess einfach möglich die korrekte Version eines Data Mining Modells zu ermitteln und dieses zu laden. Das Evaluierungsergebnis wird zu Auswertungszwecken und zur manuellen Überwachung der Funktionsweise persistiert und hat aus Prototyp Sicht keine weitere Funktion.

Dieser Prozess wird manuell angestoßen, wenn ein neues Modell zu erstellen ist. Dieser Anstoß erfolgt über den MQTT Bus mit einer Nachricht, welche den Zeitraum der Trainingsdaten für die zu erstellenden Modelle enthält. Für diese Nachricht wird am MQTT Bus ein eigenes Thema „buildmodels“ verwendet. Die Data Mining Komponente hört, für diesen Anwendungsfall, auf dieses Thema und einen validen Zeitraum in der Nachricht am MQTT Bus.

4.3.3 Vorhersageprozess

Der Anwendungsfall des Vorhersageprozesses umfasst die Erkennung des Ereignisses, wenn ein Bewohner oder eine Bewohnerin einen Raum betritt oder verlässt, sowie das anschließende Auslösen einer Vorhersage und die darauf folgende Aktion im Smart Home, wenn ein anderer Zustand als der aktuelle vorhergesagt wurde.

Der Prozess, welcher am Smartphone der Bewohner und Bewohnerinnen abläuft, ist in Abbildung 5 beschrieben. Wenn das App startet, werden zuerst die notwendigen

Berechtigungen geprüft und wenn notwendig, eingeholt. Wenn alle Voraussetzungen für einen erfolgreichen Bluetooth Suchlauf gegeben sind, wird dieser gestartet. Nach dem Suchlauf wird das Ergebnis dieses ausgewertet, wenn ein neuer Bluetooth Beacon gefunden wurde oder ein beim letzten Suchlauf noch gefundener Bluetooth Beacon, nicht mehr gefunden wurde, wird über den MQTT Bus im Smart Home eine Vorhersage für den jeweiligen Raum ausgelöst. Nach dem Auslösen oder wenn keine Veränderung am Suchergebnis festgestellt werden konnte, wird ein neuer Suchlauf gestartet. Durch dieses Vorgehen können Veränderungen an der Empfangbarkeit von Bluetooth Beacons in der App sehr zeitnah erkannt werden.

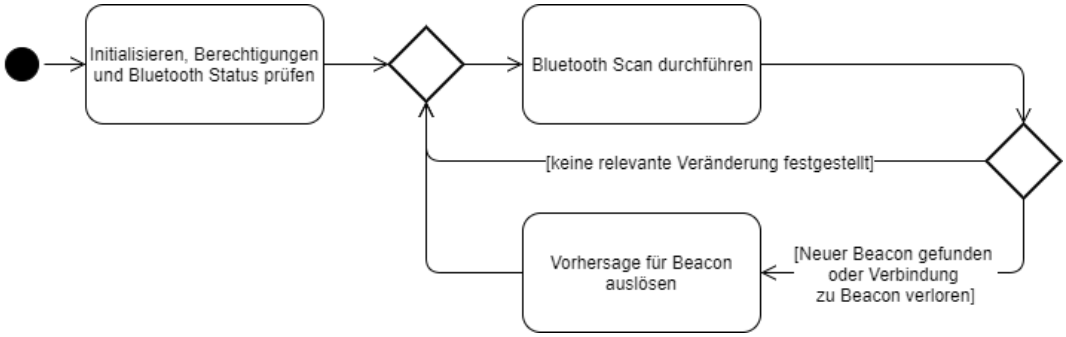

Abbildung 5: Aktivitätsdiagramm Vorhersageprozess Client Komponente

Es wird keine Abstandsberechnung zwischen Smartphone und Bluetooth Beacon durchgeführt. Sobald das Signal des Bluetooth Beacons empfangen werden kann, wird davon ausgegangen, dass sich das Smartphone im jeweiligen Raum befindet. Eine Reichweitenabgrenzung der Räume wurde durch entsprechende Platzierung der Bluetooth Beacons im Raum und durch die Konfiguration der verwendeten Sendeleistung umgesetzt. Die Signale der Bluetooth Beacons sind so nur innerhalb des Raums und auch im Bereich der Türen empfangbar.

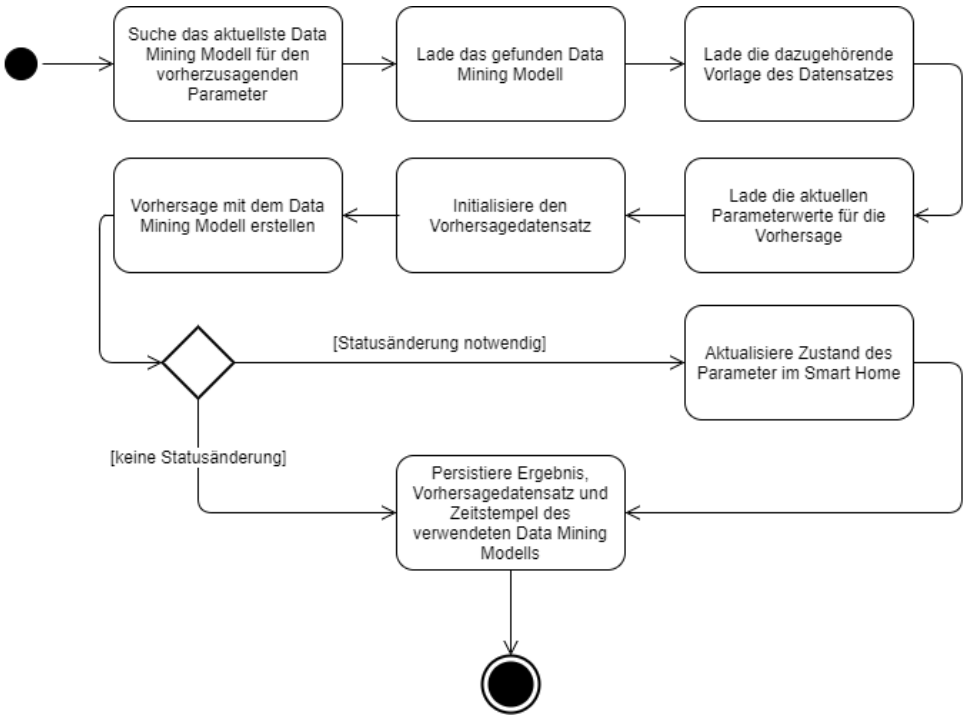

Abbildung 6: Aktivitätsdiagramm Vorhersageprozess Data Mining Komponente

Der Vorhersageprozess in der Data Mining Komponente empfängt über den MQTT Bus das Signal vom Smartphone, um eine Vorhersage zu starten. Wie in Abbildung 6 ersichtlich, wird anschließend das aktuellste Data Mining Modell für den, über den MQTTT Bus empfangen Raum gesucht. Wenn dieses Data Mining Modell in der Datenbank gefunden wurde, wird es in Weka geladen und initialisiert. Anschließend wird die dazugehörende Vorlage für einen Vorhersagedatensatz aus der Tabelle geladen und initialisiert. Von der Smart Home Komponente wird der aktuelle Zustand von allen, für die Vorhersage notwendigen, Attributen geladen. Mit diesen Werten wird anschließend der Vorhersagedatensatz befüllt. Das einzige Attribut, welches mit keinem Wert versorgt wird, ist die Beleuchtung deren Zustand vorhergesagt werden soll. Nachdem der Vorhersagedatensatz vervollständigt wurde, wird die eigentliche Vorhersage mit dem Data Mining Modell durchgeführt. Wenn der vorhergesagte Zustand der Beleuchtung ein anderer ist, als der zuvor abgerufene, wird eine Zustandsänderung über den MQTT Bus ausgelöst. Abschließend werden der für die Vorhersage verwendete Datensatz, das Vorhersageergebnis und die Version des verwendeten Data Mining Modells in einer Tabelle, für die spätere Auswertung, gespeichert.

4.4 Zusammenfassung

In diesem Kapitel wurden die Technologien für die Umsetzung des Prototyps ausgewählt, die Ergebnisse des CRISP-DM Prozesses für die im Prototyp verwendeten Data Mining Modelle, sowie die Implementierung und die Funktionsweise des Prototyps beschrieben.

Für die Technologiewahl wurde das bestehende Smart Home System und die Infrastruktur untersucht. Dabei wurden für die Integration des Prototyps in das Smart Home definiert, dass dieser als Spring Boot App in einem Docker Container in das Smart Home integriert wird. Die Data Mining Komponente kommuniziert mit Hilfe des vorhandenen MQTT Bus mit dem restlichen System. Eine Ausnahme dazu bildet die Datenextraktion, hier werden die Daten direkt aus den beiden Datenbanksystemen gelesen. Der Client des Prototyps wird in Form eines Android Apps umgesetzt und für die Präsenzerkennung, wird auf vorhandene Bluetooth Beacons gesetzt. Für das Data Mining selbst, wurde die Software Weka ausgewählt, welche in Java geschrieben ist und für schnelle Prototypen und Experimente mit Data Mining gedacht ist.

Anschließend wurde der CRISP-DM Prozess durchlaufen und die Ergebnisse dokumentiert. Im ersten Prozessschritt, dem fachlichen Verständnis, wurde erläutert, welche Faktoren aus fachlicher Sicht eine Auswirkung auf die Verwendung einer Beleuchtung haben. Dabei wurde mit Hilfe einer Literaturrecherche festgestellt, dass die Beleuchtungsnutzung von Umgebungs- und Umweltattributen, wie der Temperatur oder Tageszeit, abgeleitet werden kann. Des Weiteren wurde festgestellt, dass eine niedrige Helligkeit im Raum einen großen Einfluss auf die Wahrscheinlichkeit hat, dass jemand eine Beleuchtung aktiviert. Abschließend wurde eine Abgrenzung der an der Evaluierung beteiligten Räume, im Smart Home, durchgeführt. Zusätzlich wurde definiert, dass es notwendig ist je Beleuchtung ein eigenes Data Mining Modell zu trainieren, da jede Beleuchtung unterschiedliches Benutzungsverhalten zeigt und

unterschiedliche Informationen für diese relevant sind. Dabei wurde auch eine Einschränkung auf die Räumlichkeiten: Wohn- und Essbereich, Küche und Badezimmer vorgenommen.

Auf Basis des fachlichen Verständnisses zur Beleuchtungsnutzung, wurden im zweiten Prozessschritt, dem Verständnis der Daten, die im Smart Home vorhandenen Daten nach relevanten Informationen durchsucht und die Struktur und Qualität dieser Daten analysiert. Bei der Analyse der Datenquellen wurde festgestellt, dass eine Lückenlose Nachvollziehbarkeit aller relevanten Attribute und deren Wertveränderungen im Evaluierungszeitraum gegeben ist. Bei einer detaillierten Analyse der vorhandenen Daten, in Hinblick auf die fachlichen Anforderungen, wurden je Raum lokale, so wie globale Attribute identifiziert, welche für das Data Mining Modell relevant sind. In Summe wurden 30 Parameter identifiziert, welche für die Bildung der Datensätze verwendet werden.

Mit Hilfe dieser Schritte wurden im Prozessschritt Datenaufbereitung analysiert, in welche Struktur diese Daten transformiert werden müssen, damit sie anschließend für das Data Mining Modell verwendet werden können. Dabei wurde definiert, dass je Minute im relevanten Zeitraum, eine Instanz mit den, zu diesem Zeitpunkt gültigen Wert aller Attribute, gebaut wird. Der dadurch entstehende Datensatz enthält zu jeder Minute im enthaltenen Zeitraum eine Instanz, mit den jeweiligen Werten zu jedem der 30 identifizierten Attribute. Dieser Datensatz bildet die Basis für alle zu trainierenden Data Mining Modelle. Je nach Beleuchtung wird jedoch eine unterschiedliche Teilmenge an Attributen für den Trainingsdatensatz verwendet.

Auf Basis dieses Datensatzes wurde das Training und die Evaluierung der initialen Data Mining Modelle begonnen. Zu Beginn wurde der Algorithmus J48 für das Training ausgewählt und ein erster Trainingslauf für alle sechs Data Mining Modelle durchgeführt. Dabei wurde festgestellt, dass es, aufgrund der ungleichen Verteilung der Klassen, notwendig ist, die Vorhersagewerte für die Klassen „Beleuchtung aus“ und „Beleuchtung ein“ getrennt zu behandeln. Aufgrund der Qualität der Vorhersagewerte wurde im CRISP-DM Prozess eine Iteration durchgeführt und die Attributauswahl für die Datensätze, sowie die Parametrisierung des Algorithmus optimiert. Durch diese Verbesserungsmaßnahmen konnte die Vorhersagequalität deutlich gesteigert werden. In einem zweiten Durchlauf konnten somit Vorhersagewerte von über 83% korrekter Vorhersagen für fünf der sechs Data Mining Modelle erzielt werden. Beim sechsten Data Mining Modell war keine Steigerung über 45,6% möglich.

Abschließend wurde die Implementierung des Prototyps beschrieben. Dieser besteht aus zwei Komponenten, welche in das bestehende System des Smart Home integriert worden sind. Die Hauptkomponente ist eine Server Komponente, welche für die Datenextraktion, das Training von Data Mining Modellen und das Treffen von Vorhersagen mit den trainierten Data Mining Modellen, zuständig ist. Die zweite Komponente ist eine Client Komponente, welche in Form einer Android App, umgesetzt wurde. Diese Komponente ist zuständig für das Erkennen von Ereignissen, die eine neue Vorhersage des Zustands einer Beleuchtung erfordern. Die Detektion solcher Ereignisse wurde mit Hilfe von Bluetooth Beacons umgesetzt.

Im folgenden Kapitel wird auf die Evaluierung des, in diesem Kapitel vorgestellten Prototyps, zur Beantwortung der Forschungsfrage eingegangen und die Fallstudie strukturiert und geplant.

5 EVALUIERUNG

Dieses Kapitel beschreibt den Aufbau und die Vorgehensweise zur Evaluierung der Data Mining Modelle und des zuvor in Kapitel 4 beschriebenen Prototyps. Zu Beginn wird mit Hilfe einer Literaturrecherche zu ähnlichen Arbeiten im Bereich des Smart Home die Vorgehensweise zur Planung, Durchführung und Auswertung der Fallstudie erörtert. Anschließend wird der Aufbau des abschließenden Interviews mit den teilnehmenden Bewohnern und Bewohnerinnen beschrieben. Am Ende des Kapitels wird der konkrete Ablauf der Evaluierung geplant und die Durchführung mit den recherchierten Werkzeugen und Methoden beschrieben und dokumentiert.

5.1 Evaluierung von Data Mining Modellen im Smart Home

Auf die grundsätzlichen Möglichkeiten, Werkzeuge und das Vorgehen zur Evaluierung von Data Mining Modellen, wurde bereits in Abschnitt 2.3.2 eingegangen. Für die Evaluierung in der Fallstudie der vorliegenden Arbeit, im Bereich des Smart Homes und zur Beantwortung der Forschungsfrage wurden zwei Themen identifiziert, die detaillierter aufbereitet werden mussten. Zuerst wurde, durch eine Literaturrecherche erörtert, wie eine Evaluierung der Vorhersagequalität von Data Mining Modellen im Bereich des Smart Homes durchgeführt werden kann. Anschließend wurde ermittelt, wie, für den Anwendungsfall der vorliegenden Arbeit, die Lernfähigkeit dieser Data Mining Modelle evaluiert werden kann.

5.1.1 Evaluierung der Vorhersagequalität

Dieser Abschnitt befasst sich damit, wie im Umfeld von Smart Homes eine praktische Evaluierung von Data Mining Modellen, sowie die Evaluierung der Lernfähigkeit dieser, an sich ändernde Umweltbedingungen, durchgeführt werden kann.

Als Ergänzung zur in Abschnitt 2.3.2 beschriebenen Konfusionsmatrix beschreiben Kotu und Deshpande (2015) auch die Grenzwertoptimierungskurven (Englisch: ROC curves) als wichtiges Werkzeuge, um die Vorhersagequalität eines Data Mining Modells zu evaluieren. Bei der Grenzwertoptimierungskurve werden die als richtig positiv klassifizierten Instanzen in einem zweidimensionalen Koordinatensystem auf der einen Achse, gegen die falsch positiv klassifizierten Instanzen auf der anderen Achse dargestellt. Auf diese Art und Weise ist ein einfacher grafischer Vergleich von Data Mining Modellen möglich.

In den Daten des Anwendungsfalls in der vorliegenden Arbeit, gibt es ein sehr großes Ungleichgewicht bei der Verteilung der vorherzusagenden Klassen. Es gibt eine sehr große Anzahl von Instanzen, welche die Klasse „Beleuchtung aus“ repräsentieren und dazu aber nur

eine verhältnismäßig kleine Anzahl von Instanzen welche die Klasse „Beleuchtung ein“ repräsentieren. Zum Beispiel sind für das Attribut der Beleuchtung Spielecke, im in Abschnitt 4.2.4 verwendeten Datensatz, 82346 Instanzen mit der Klasse „Beleuchtung aus“ aber nur 2614 Instanzen mit der Klasse „Beleuchtung ein“ vorhanden. Dieses Ungleichgewicht erschwert die Evaluierung des Data Mining Modells. Prati, Batista und Monard (2009) haben sich mit dem Problem der ungleichen Klassenverteilung beschäftigt und haben das Problem der Evaluierung analysiert. Sie beschreiben, dass weit verbreitete Indikatoren für die Evaluierung von Data Mining Modellen, wie der gesamten Fehlerrate oder die Präzision bei solchen Modellen, oft sehr falsche Ergebnisse liefern. Wenn das Data Mining Modell für die stärker vertretenen Klassen sehr gute Ergebnisse liefert, gehen schlechte Vorhersagewerte für die weniger vertretenen Klassen, in diesen Indikatoren, aufgrund der ungleichen Verteilung, unter. Dieses Verhalten wurde in der vorliegenden Arbeit in Abschnitt 4.2.3 identifiziert und dadurch gelöst, in dem die Vorhersagewerte der einzelnen Klassen getrennt betrachtet wurden. Prati, Batista und Monard (2009) schlagen diese, in Abschnitt 4.2.3 bereits durchgeführte, getrennte Betrachtung der Vorhersagewerte mit Ergänzung durch Grenzwertoptimierungskurven für die Evaluierung von Zwei-Klassen Data Mining Modellen vor. Für die Evaluierung der Data Mining Modelle im Prototyp der vorliegenden Arbeit, wurde auf die zuvor beschriebenen Werkzeuge zurückgegriffen. Es wurden die Vorhersagewerte der beiden Klassen getrennt mit Konfusionsmatrizen ausgewertet und mit Grenzwertoptimierungskurven für eine einfachere Analyse dargestellt.

Zur Evaluierung von Data Mining Modellen im Kontext von Smart Homes, wurde eine Literaturrecherche auf Google Scholar mit dem Schwerpunkt auf wissenschaftliche Artikel und Konferenzpublikationen zu Fallstudien aus diesem Bereich durchgeführt. Auf für die Evaluierung der vorliegenden Arbeit relevante Arbeiten, welche durch die Literaturrecherche gefunden wurden, wird in weiterer Folge eingegangen. Als relevant wurden dabei gefundene Arbeiten eingestuft, in denen Data Mining Verfahren für Klassifizierungsprobleme in einem Smart Home zur Erkennung von Mustern oder zur Automatisierung von Aufgaben evaluiert wurden.

Cook, Crandall, Thomas und Krishnan (2013) haben in ihrem CASA Projekt eine Fallstudie in mehreren Smart Homes durchgeführt, um komplexe Aktivitäten in diesen erkennen zu können. Dabei wurden in normalen Wohnungen vernetzte Sensoren installiert, welche das Verhalten der Bewohner und Bewohnerinnen aufgezeichnet haben. Von den Sensoren wurde jede Veränderung des Zustandes gesammelt und in den Datenbestand aufgenommen. In Ihrem Fall konnte auf keine bestehende Datenbasis zurückgegriffen werden, da ein normales Smart Home komplexe Aktivitäten nicht erkennen und aufzeichnen kann. Um eine Datenbasis für das Training des Data Mining Modells zu schaffen, wurden ein Monat lang Daten aufgezeichnet und anschließend manuell zu komplexen Aktivitäten wie Kochen oder Schlafen zugewiesen. Diese, manuell klassifizierten Daten, über den Zeitraum eines Monates, wurden anschließend für das Training des Data Mining Modells verwendet. Anschließend wurde das Data Mining Modell, ohne neues Trainieren, in mehreren Wohnungen eingesetzt und Klassifizierungen zu den neu gesammelten Daten vorgenommen. Bei diesen Klassifizierungen konnte im Durschnitt eine Präzision von 84% der klassifizierten Aktivitäten erreicht werden. Zusammengefasst wurde hier

ein vortrainiertes Data Mining Modell eingesetzt und am Ende einmalig die Gesamtpräzision des zuvor trainierten Modells mit den, in den Wohnungen neu gesammelten Daten ausgewertet.

Tapia, Intille und Larson (2004) haben eine ähnliche Studie durchgeführt. Ein Prototyp in einem Smart Home wurde hier verwendet, um die, für die Evaluierung des Data Mining Modells benötigten Daten, über einen Zeitraum von 14 Tagen, aufzuzeichnen und zu klassifizieren. Dabei wurden alle Zustandsveränderungen der Sensoren aufgezeichnet. In dieser Studie haben die Bewohner und Bewohnerinnen des Smart Homes mittels einem PDA selbst laufend klassifiziert, welche Aktivität sie gerade ausüben. Auf diese Art ist ein Datensatz aufgebaut worden, welcher anschließend in einen Trainings- und einen Evaluierungsdatensatz aufgeteilt wurde. In dieser Arbeit wurde der gesamte gesammelte Datenbestand für die Evaluierung verwendet, um ein Data Mining Modell trainieren und evaluieren zu können.

In der Arbeit von Youngblood, Heiermann, Holder und Cook (2015) wurde ein Data Mining Modell zur Automatisierung der Beleuchtung in einem Labor Smart Home trainiert und evaluiert. Sie haben Data Mining Modelle mit vorhandenen Daten trainiert und im Tagesbetrieb eines Smart Homes, Vorhersagen zum Zustand von Beleuchtungen durchgeführt. Dabei wurde die Anzahl von manuellen Interaktionen der Bewohner und Bewohnerinnen im Smart Home aufgezeichnet, um auf diese Weise feststellen zu können, ob durch den Einsatz des Data Mining Modells eine Verringerung der manuellen Interaktion mit dem Smart Home erreicht werden kann. Das Data Mining Modell wurde mit einer Woche an Trainingsdaten mit 1400 Instanzen pro Tag trainiert. Das entspricht, wie in der vorliegenden Arbeit, einer Instanz pro Minute. Dieses Vorgehen wurde mit unterschiedlichen Data Mining Verfahren durchgeführt, um diese vergleichen zu können. Zusammengefasst verfolgt diese Arbeit ein ähnliches Ziel, wie die Fallstudie der vorliegenden Arbeit. Durch das Data Mining Modell wird direkt in die Automatisierung der Haussteuerung eingegriffen und dem Benutzer notwendige Interaktionen mit dem Smart Home abgenommen.

Spring, Cook, Weeks, Dahmen und La Fleur (2017) haben in ihrer Arbeit für die Analyse von Zeitseriendaten, unter anderem, Klassifizierungsmodelle eingesetzt. Sie haben nicht ein großes Modell trainiert und dieses anschließend eingesetzt, sondern sie haben in ihren Zeitseriendaten dasselbe Data Mining Modell mit den Daten aus unterschiedlichen Zeiträumen trainiert um auf diese Art und Weise, Unterschiede zwischen den Zeiträumen erkennen zu können. Sie zeigen damit, dass es sinnvoll sein kann, dasselbe Data Mining Modell mit unterschiedlichen Datenbeständen von Daten mit Zeitbezug zu trainieren, um Unterschiede durch neue Zeitbereiche erkennen zu können.

Aus dieser Literaturrecherche konnte abgeleitet werden, dass eine Evaluierung von Data Mining Modellen in einem Smart Home mit echten gesammelten und klassifizierten Daten erfolgen sollte. Dabei werden über einen definierten Zeitraum sämtliche relevante Daten und Ereignisse gesammelt, welche anschließend entweder automatisiert oder manuell klassifiziert werden. Diese Daten werden zur Erstellung der initialen Data Mining Modelle verwendet. Anschließend werden die trainierten Data Mining Modelle in das Smart Home integriert, wobei laufend Vorhersagen durch eintretende Ereignisse getroffen werden. Auch diese Vorhersagen müssen, automatisiert oder manuell klassifiziert werden. Diese Klassifizierung wird anschließend der

Vorhersage der Data Mining Modelle gegenübergestellt. Durch diese Gegenüberstellung ist die Ermittlung der Konfusionsmatrix und damit eine Aussage über die Vorhersagequalität der Data Mining Modelle möglich. Die Konfusionsmatrix wird hier, wie auch in Abschnitt 4.2.4, in Prozentzahlen der korrekten und falschen Vorhersagen je Klasse ausgewertet. Zusätzlich sollte bei dieser Evaluierung auch die Wirkung im Smart Home beobachtet werden. Wie entwickeln sich die manuellen Korrekturen der falschen Vorhersagen und wie hilft das Data Mining Modell manuelle Interaktion mit dem Smart Home zu reduzieren.

5.1.2 Evaluierung der Lernfähigkeit

Um die Lernfähigkeit des Data Mining Modells evaluieren zu können, muss während der Durchführung der Fallstudie, eine Veränderung im Benützungsverhalten der vorherzusagenden Beleuchtung stattfinden. Mit der Veränderung des Benutzungsverhaltens geht eine Veränderung der Datenpopulation einher, indem sich Attribute und Instanzen in der Datenbasis verändern. Hulten, Spencer und Domingos (2001) beschreiben, dass solche Veränderungen, wenn sie groß genug sind, zur Verschlechterung der Vorhersagewerte und damit zur Erhöhung der Fehlerrate von Data Mining Modellen führen können. Sie beschreiben eine auf sehr große Datenmengen ausgerichtete Technik für ein kontinuierliches Aktualisieren von Data Mining Modellen. Bei ihrer Evaluierung haben sie die Fehlerraten von unterschiedlichen Versionen der Data Mining Modelle für die Vorhersage mit denselben Testdaten verglichen. Sie haben so die Vorhersagequalität von einem alten Data Mining Modell und die Vorhersagequalität von einem laufend aktualisierten Data Mining Modell, mit aktuellen Testdaten verglichen und konnten auf diese Weise die Verbesserung durch das kontinuierliche aktualisieren feststellen. Ziel der Evaluierung in der vorliegenden Arbeit ist es, herauszufinden, ob diese Verschlechterung durch aktualisieren der Data Mining Modelle auch im vorliegenden Anwendungsfall, im Bereich des Smart Homes entgegengewirkt werden kann und so die Vorhersagequalität verbessert werden kann.

Für die Evaluierung in der vorliegenden Arbeit soll die natürliche, im Jahresverlauf stattfindende Schwankung der Sonnenzeiten und die damit verbundene Veränderung im Benützungsverhalten von Beleuchtungen, die Grundlage für die Evaluierung bilden. Am Standort des Smart Homes, in welchem die Fallstudie durchgeführt wurde, verringert sich laut der Berechnung von timeanddate.de (2017), die Zeit mit Tageslicht im Zeitraum von Anfang Juli bis Ende November von 15 Stunden und 49 Minuten auf 8 Stunden und 51 Minuten. Der Zeitraum am Tag mit Tageslicht verringert sich in diesem Zeitraum damit um 44%. Der Zeitpunkt des Sonnenuntergangs verschiebt sich dabei von 20:54 auf 16:14. Hunt (1979) erkannte, wie bereits in Abschnitt 4.2.1 beschrieben, einen Zusammenhang zwischen der Helligkeit im Raum und der Wahrscheinlichkeit, dass eine Beleuchtung aktiviert wird. Durch den, in diesem Zeitraum, immer früher beginnenden Sonnenuntergang, sinkt die Helligkeit im Raum bereits zu einer früheren Uhrzeit. Damit steigt die Wahrscheinlichkeit, dass die Bewohner und Bewohnerinnen bereits zu einer früheren Uhrzeit den Bedarf an künstlicher Beleuchtung haben. Eine Auswertung der im Jahr 2016 erfassten Daten zu den in der Fallstudie behandelten Beleuchtungen im Smart Home haben ergeben, dass sich die Anzahl der

Beleuchtungsaktivierungen in diesem Zeitraum erhöht. Wie in Abbildung 7 ersichtlich ist, hat sich bei allen betrachteten Beleuchtungen die Anzahl der Aktivierungen pro Monat in diesem Zeitraum erhöht. Besonders auffallend ist der Anstieg der Beleuchtungsaktivierung im Badezimmer, aber auch die Beleuchtungen in Küche, am Esstisch, beim Fernseher und in der Spielecke wurden sichtlich häufiger verwendet.

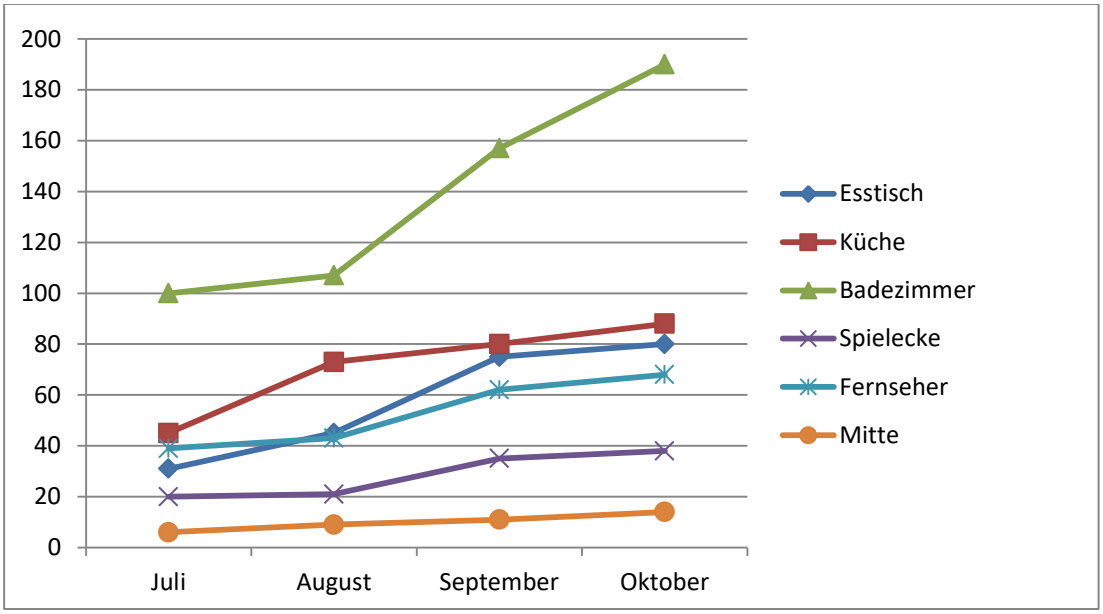

Abbildung 7: Anzahl der Beleuchtungsaktivierungen im Zeitraum Juli 2016 bis Oktober 2016

Diese Steigerung der Beleuchtungsnutzung durch die abnehmende Sonnenzeit zwischen den Monaten Juli und November ist ein Effekt mit dem, die zuvor beschriebene Lernfähigkeit, durch das Aktualisieren der Data Mining Modelle im Prototyp evaluiert werden kann.

Kelly Hand und Adams (1999) beschäftigten sich in ihrer Arbeit mit den Auswirkungen von sich kontinuierlich verändernden Datenpopulation im Zusammenhang mit der Fehlerrate von Data Mining Modellen. Sie haben festgestellt, dass bei einer sich kontinuierlich verändernden Datenpopulation, sich auch die Fehlerrate kontinuierlich verschlechtert. Von einer sich kontinuierlich im Jahresverlauf ändernden Datenpopulation kann auch in der Fallstudie der vorliegenden Arbeit ausgegangen werden, da die Veränderung der Beleuchtungsbenutzung, durch den sich ändernden Tageslichtzyklus, kontinuierlich auf die Datenpopulation wirkt. Kelly Hand und Adams (1999) haben, wie auch Hulten, Spencer und Domingos (2001) gezeigt, dass man durch das Treffen von Vorhersagen mit Testdaten aus der ursprünglichen und der neuen Datenpopulation die Unterschiede in den Vorhersagewerten, durch eine Steigerung der Fehlerrate erkennen kann. Es ist auch möglich, bei einer kontinuierlichen Vorhersage von Zeitseriendaten, den Trend der Veränderung gut nachzuvollziehen und zu erkennen, ab wann sich eine Datenpopulation zu weit verändert.

Aus diesen Recherchen und Ergebnissen ergeben sich neue Erkenntnisse für die Evaluierung der Data Mining Modelle in der Fallstudie der vorliegenden Arbeit. Es muss mehr als eine Evaluierung durchgeführt werden. Zuerst muss eine Evaluierung der Data Mining Modelle mit Testdaten aus der Population des initialen Trainingsstands der Data Mining Modelle durchgeführt werden, um die Vorhersagewerte mit der initialen Datenpopulation zu ermitteln.

Anschließend müssen die Data Mining Modelle mit neueren Daten aus der sich veränderten Datenpopulation evaluiert werden, um die Vorhersagewerte mit Daten aus der neueren Datenpopulation zu ermitteln. Diese Evaluierungen müssen sowohl mit der initialen, als auch mit der aktualisierten Version des Data Mining Modells durchgeführt werden. So werden sowohl Daten aus der ursprünglichen, als auch aus der neuen Datenpopulation mit dem ursprünglichen und dem kontinuierlich trainierten Data Mining Modell evaluiert. Auf Basis dieser Vorhersagewerte können Rückschlüsse für die Interpretation der Lernfähigkeit abgeleitet werden. Bei näherer Betrachtung der Auswertung zur Beleuchtungsaktivierung im Jahr 2016 wurde festgestellt, dass bei der Auswahl der Zeitfenster für die Datensätze darauf geachtet werden muss, die zeitlichen Abstände nicht zu klein zu wählen. Wenn der zeitliche Abstand zwischen initialer Datenpopulation und veränderter Datenpopulation zu klein ist, wird im Anwendungsfall der vorliegenden Arbeit, auch die Veränderung der Datenpopulation zu klein sein, um Auswirkungen zu zeigen.

5.2 Fallstudie

In diesem Abschnitt wird der Aufbau der Fallstudie, welche in der vorliegenden Arbeit durchgeführt wurde, beschrieben. Zuerst wird mit Hilfe von Literatur die Fallstudie als Methodik, sowie deren grundsätzliche Struktur und Ablauf beschrieben. Anschließend wird dieses Vorgehen für die vorliegende Arbeit umgesetzt und das Design und der Ablauf der durchgeführten Fallstudie beschrieben.

5.2.1 Definition und Ablauf der Fallstudie

Yin (2013) definiert eine Fallstudie als empirische Untersuchung im natürlichen Anwendungsumfeld eines Phänomens, bei dem die Grenzen zwischen Phänomen und Umwelt nicht klar sind. Die wichtigen Aspekte bei Fallstudien sind, dass sie in realen Anwendungen durchgeführt werden und so Beobachtungen und Datenerfassung des Untersuchten Phänomens im realen Auftreten möglich ist. Er beschreibt auch, dass Fallstudien eine gute Wahl sein können, wenn wenig Kontrolle über die zu beobachtenden Ereignisse besteht und diese in einem realen Anwendungsumfeld stattfinden.

Stake (1995) definiert eine Fallstudie als Untersuchung der Besonderheit und Komplexität eines Falles unter bestimmten Gegebenheiten mit dem Ziel, Aktivitäten innerhalb des Falles besser zu verstehen. Er definiert den Fall selbst als integriertes System, in welchem Aktivitäten unter bestimmten Gegebenheiten ablaufen. Dieses System muss dabei ein spezifisches und komplexes System sein.

Yin (2013) beschreibt, dass Fallstudien für deskriptive, explanative oder explorative Anwendungsfälle angewandt werden können. Unter deskriptiver Anwendung versteht Yin (2013) die Erfassung und Beschreibung eines Phänomens, welches zuvor bereits in der Theorie beschrieben und erarbeitet wurde. Unter explanativer Anwendung versteht Yin (2013) die Erklärung und Begründung von zuvor in der Theorie aufgestellten Aussagen. Unter

explorativer Anwendung versteht Yin (2013) die Untersuchung und Beobachtung eines Phänomens, um die aufgestellten Hypothesen und Theorien zu verfeinern.

Laut Yin (2013) kann jede Fallstudie, sowohl als Einzelfallstudie, bei der ein einzelner Fall untersucht wird, als auch als mehrfache Fallstudie, bei der mehrere Fälle parallel oder sequentiell untersucht werden, geplant und durchgeführt werden. Ob eine Einzelfallstudie oder eine mehrfache Fallstudie durchgeführt werden sollen, lässt sich nicht generell beantworten. Mehrfache Fallstudien sind mit einem sehr hohen Ressourcenaufwand verbunden und bedingen, dass mehr als ein Fall vorliegt. Er empfiehlt jedoch, wenn es möglich ist, eine mehrfache Fallstudie, mit zumindest zwei Fällen, durchzuführen.

Ein großer Vorteil in Fallstudien ist die Triangulation. Mit diesem Begriff ist im Kontext von Fallstudien, die Kombination von Datenerhebungsmethoden einzusetzen, um einen möglichst vollständigen Überblick über den Fall und das zu untersuchende Ereignis zu bekommen. Dabei können sowohl qualitative, als auch quantitative Daten kombiniert werden. (Stake, 1995)

Für die vorliegende Arbeit wird eine einfache Fallstudie durchgeführt. Da nur ein Smart Home für die Evaluierung zur Verfügung steht, ist es nicht sinnvoll, eine mehrfache Fallstudie durchzuführen. Für den konkreten Anwendungsfall wird auf eine deskriptive Anwendung der Fallstudie zurückgegriffen. Mit der erfassten Beobachtung sollen die zuvor in der Theorie aufgestellten Hypothesen, durch die Beobachtung beschrieben und bestätigt werden.

Rowley (2002) fasst die Verwendung von Fallstudien in der Forschung zusammen und fasst dabei die Ergebnisse von mehreren Arbeiten zusammen. Unter anderem sind hier auch die Arbeiten von Yin (1994) und Stake (1995) gelistet. Rowley (2002) hat den Ansatz und die Komponenten des Designs von Yin (1994) übernommen und um zusätzliche Informationen ergänzt. Eisenhardt (1989) beschäftigte sich in ihrer Arbeit mit der Bildung von Theorien durch Fallstudien. Dabei hat sie einen Prozess für die Vorgehensweise zur Theorienbildung aus Fallstudien entwickelt. Die Bestandteile dieses Prozesses decken sich mit den Komponenten von Yin (1994). Das Design einer Fallstudie muss, nach Yin (1994), zumindest über die folgenden Komponenten verfügen:

- Forschungsfrage
- Hypothesen
- Untersuchte Objekte
- Beziehung zwischen Daten und Hypothesen
- Kriterien um die Ergebnisse interpretieren zu können

Für die Durchführung einer Fallstudie ist es unumgänglich, Forschungsfragen zu definieren, welche die Richtung und das Ziel der Fallstudie vorgeben. Aus diesen Forschungsfragen müssen Hypothesen abgeleitet werden, nur durch diese Hypothesen wird klar definiert, in welche Richtung gesucht werden soll. Für die Durchführung der Fallstudie ist es notwendig, die zu untersuchenden Objekte genau zu spezifizieren, damit die Datenerhebung richtig strukturiert werden kann und die richtigen Objekte beobachtet werden. Für eine korrekte Datenerhebung ist es außerdem von Relevanz, dass eine Beziehung zwischen den erhobenen Daten und den

aufgestellten Hypothesen definiert wird. Welche Daten sind notwendig, um die Hypothesen belegen oder wiederlegen zu können. Zum Schluss müssen Kriterien definiert werden, wie die erhobenen Daten zu interpretieren sind. (Yin, 2013)

Noor (2008) beschreibt den Prozess zur Durchführung von Fallstudien. Als ersten Schritt dieses Prozesses sieht er folgende Punkte. Das Design der Fallstudie, die Formulierung der Forschungsfrage und die Ableitung der Hypothesen. In diesem Schritt sieht er auch die Auswahl der zu untersuchenden Fälle. Im zweiten Schritt sollen die Fallstudien durchgeführt und die Daten erhoben werden. Anschließend werden die erhobenen Daten, je untersuchten Fall, analysiert. Auf Basis dieser Einzelanalysen kann anschließend eine Analyse über alle untersuchten Fälle durchgeführt werden. Die durchgeführten Analysen bilden die Basis für die Schlussfolgerung und Anpassung der aufgestellten Theorie.

Für die vorliegende Arbeit wurde eine Fallstudie mit den Komponenten nach Yin (2013) und dem Ablaufprozess nach Noor (2008) entworfen. Durch den bereits vorliegenden Fall und der Tatsache, dass eine einfache Fallstudie durchgeführt wurde, wurden die Prozessschritte der Fallauswahl, sowie die fallübergreifende Analyse, nicht berücksichtigt. Das Design und die Planung dieser Fallstudie sind im nächsten Abschnitt im Detail beschrieben.

5.2.2 Entwurf der Fallstudie für die vorliegende Arbeit

Als ersten Schritt wurde der Fall der einfachen Fallstudie definiert: Das bestehende Smart Home mit dem integrierten Prototyp zur Beleuchtungssteuerung, sowie die Bewohner und Bewohnerinnen wurden als Fall der vorliegenden Fallstudie definiert. Im bestehenden Smart Home befinden sich über 100 installierte Sensoren und Aktoren welche in eine zentrale Smart Home Komponente integriert sind. Diese Komponente wertet alle Sensordaten aus und übernimmt die Steuerung der Aktoren auf Basis von einem Regelwerk. Diese bestehende Logik wurde um den Prototyp zur vollständigen Automatisierung der Beleuchtungssteuerung ergänzt. Alle Werte und Zustände, sowie deren Historie, werden persistiert. Das Smart Home wird von einem männlichen Bewohner (27 Jahre), einer weiblichen Bewohnerin (28 Jahre), sowie von einem Kleinkind (1 Jahr) bewohnt. Der männliche Bewohner geht einer Vollzeitbeschäftigung nach und verbringt somit primär seine Freizeit im Smart Home. Die weibliche Bewohnerin befindet sich aktuell in Karenz zu Hause beim Kleinkind und verbringt somit den Großteil der Zeit im Smart Home. Die beiden erwachsenen Bewohner sind mit einem Smartphone und der entsprechenden App für die Prototypen der vorliegenden Arbeit ausgestattet, welche Vorhersagen für die Beleuchtungen durchführt und Statusveränderungen auslöst. Die Bewohner sind angewiesen das Smartphone ständig bei sich zu führen und die App aktiv zu haben. Die Bewohner führen ein Protokoll zu Anomalien bei den von den Prototypen durchgeführten Änderungen am Zustand der Beleuchtung, sowie zu notwendigen manuellen Beleuchtungsaktivierungen. Die Bewohner sind auch dazu angehalten, ein stichwortartiges Tagebuch über den allgemeinen Eindruck der Vorhersageleistung des Prototyps zu erstellen. Dieses Tagebuch soll die Datenerhebung in der Fallstudie durch qualitative Informationen ergänzen.

Anschließend werden der Aufbau und die Durchführung der Fallstudie für die vorliegende Arbeit entworfen und strukturiert. Die Grundlage dafür bilden die zuvor diskutierten Komponenten nach Yin (2013) und der Prozess nach Noor (2008). In Folge wird das Design der Fallstudie entsprechend der Komponenten nach Yin (2013) beschrieben:

Die beiden ersten Komponenten des Forschungsdesigns nach Yin (2013), die Forschungsfrage und die Hypothesen wurden bereits in Abschnitt 1.2 behandelt und definiert. Diese Definitionen werden für die Durchführung der Fallstudie verwendet.

Als zu untersuchende Objekte im Fall wurden, der erstellte Prototyp für die Beleuchtungssteuerung, die Interaktionen der Bewohner mit dem Smart Home, sowie alle vom Smart Home gesammelten Daten definiert. Außerdem müssen die in dem Protokoll und manuellen Tagebuch geführten Daten erhoben werden, aus diesem Grund wurden auch diese beiden Protokolle als zu untersuchende Objekte deklariert. Eine Beobachtung dieser Objekte und Erhebung der relevanten Daten zu diesen Objekten ist für die Auswertung der Fallstudie notwendig. Konkret werden alle Beleuchtungsaktivierungen und Beleuchtungsdeaktivierungen, welche ohne oder mit Einwirkung des Prototyps vorgenommen wurden, für die Auswertung aufgezeichnet. Alle im Evaluierungszeitraum entstehenden Instanzen der Trainingsdaten, unabhängig davon, ob sie für das Training verwendet wurden oder nicht, werden für die Auswertung aufgezeichnet. Alle Versionen des Prototyps, sowie die Evaluierungswerte dazu und alle eingehenden Vorhersageanfragen, werden für die Auswertung aufgezeichnet. Die Inhalte der manuell geführten Protokolle und Tagebücher werden für die Auswertung erhoben.

Aus den gesammelten Daten zu manuellen Beleuchtungsänderungen, den automatisch durch den Prototyp durchgeführten Beleuchtungsänderungen und den durchgeführten Vorhersagen, kann ein Testdatensatz für die Evaluierung erstellt werden. Für die automatische Klassifizierung wird die folgende Regel angewandt: Wird in einem Zeitraum von 30 Sekunden, um den Vorhersagezeitraum der Status einer Beleuchtung nicht manuell verändert, hat das Data Mining Modell eine korrekte Vorhersage getroffen. Mit dieser Logik können aber nicht alle Fälle erfasst werden, wodurch anschließend noch die, in den Protokollen angeführten Anomalien ausgewertet und berücksichtigt werden müssen. Dieser erfasste Testdatensatz kann dann gegen das Modell evaluiert werden, um die Vorhersagequalität zu ermitteln. Die ermittelte Vorhersagequalität stellt anschließend den Bezug zur Hypothese her. Ergänzend zu diesen Werten wird auch das Tagebuch ausgewertet, um eine qualitative Einschätzung zum Lernverhalten des Prototyps zu bekommen.

Die Fallstudie muss im gleichen Fall mit zwei unterschiedlichen Versionen der Data Mining Modelle durchlaufen werden. Der Vergleich der Vorhersagewerte dieser beiden Durchläufe ermöglicht Rückschlüsse auf die Gültigkeit der Hypothesen.

Damit sind alle relevanten Komponenten der Fallstudie beschrieben und der erste Schritt im Prozess zur Durchführung einer Fallstudie nach Noor (2008) ist abgeschlossen: Für den zweiten Schritt dieses Prozesses der Durchführung der Fallstudie wurde der folgend beschriebene Plan aufgestellt.

Als Vorphase zur eigentlichen Fallstudie, werden im Zeitraum von Anfang Juli 2017 bis zum Ende der Evaluierung im Smart Home, alle Statusveränderungen von Beleuchtungen, sowie

alle weiteren, für den Trainingsdatensatz aus 4.2.3 relevanten Sensordaten, aufgezeichnet. Aus diesen Daten werden anschließend Trainingsdatensätze für die Fallstudie erstellt. Die Fallstudie muss mit zwei Durchläufen, zu je einer Woche, im selben Smart Home durchgeführt werden. Diese beiden Durchläufe wurden in den ersten beiden Novemberwochen geplant. Dieser Abstand von mehreren Monaten, zwischen dem Beginn der Datensammlung und der Durchführung der Evaluierung, ist notwendig, damit der in Abschnitt 5.1.2 definierte Effekt, Auswirkungen auf das Benützungsverhalten der Beleuchtungen zeigt. In der ersten Woche wird im Prototyp ein Data Mining Modell verwendet, welches mit einem Trainingsdatensatz erstellt wurde, das den Zeitraum von Anfang Juli bis Ende August enthält. Bei diesem Durchlauf werden die Vorhersagewerte für ein nur initial trainiertes und nie aktualisiertes Data Mining Modell evaluiert. Für den zweiten Durchlauf wird ein Data Mining Modell verwendet, welches mit einem Trainingsdatensatz erstellt wurde, das den Zeitraum von Anfang Juli bis Ende Oktober enthält. Bei diesem Durchlauf werden die Vorhersagewerte für ein kontinuierlich aktualisiertes Data Mining Modell evaluiert. Dieses Data Mining Modell spiegelt den Stand eines Data Mining Modells wieder welches im Zeitraum von Beginn der Datenaufzeichnung bis zur Vorhersage wöchentlich neu aufgebaut wurde.

Nach Abschluss der beiden Durchläufe können die Vorhersagewerte ausgewertet und verglichen werden. Durch die Auswertung und den Vergleich der Vorhersagewerte ist es möglich, die Hypothesen zu belegen, oder zu verwerfen und die Forschungsfrage zu beantworten.

5.3 Aufbau des abschließenden Interviews

Abschließend soll ein Interview des Bewohners und der Bewohnerin durchgeführt werden, um qualitative Informationen zum evaluierten Prototyp zu erheben. Dabei soll der subjektive Eindruck der Vorhersageleistung allgemein, sowie der Unterschied zwischen den beiden Evaluierungen erhoben werden. Zusätzlich soll auch der Nutzen eines solchen Vorhersage- und Entscheidungssystems, aus Sicht der Bewohner und Bewohnerinnen des Smart Homes, erhoben werden.

Als Form des Interviews wurde eine Fokusgruppe gewählt. Diese Form des Interviews ist nach Flick (2010) als teilstandardisiertes Interview einzustufen. Laut Krueger und Casey (2014) sind Fokusgruppen dafür gedacht, qualitatives Empfinden und Meinungen von den Teilnehmern und Teilnehmerinnen zu erheben und sie sollen dazu beitragen, die Meinungen und den Gedanken besser zu verstehen. Laut Krueger und Casey (2014) besteht eine Fokusgruppe aus einer kleinen Zahl von Personen, welche über dasselbe Thema diskutieren und dabei qualitative Daten bereitstellen, welche dabei helfen das Thema besser zu verstehen.

Um den Einsatz von Fokusgruppen im Bereich des Smart Homes zu prüfen und um die Umsetzung, in der vorliegenden Arbeit, an solchen Beispielen orientieren zu können, wurde eine Literaturrecherche nach Zeitschriftenartikeln zu den Stichworten „smart home“ und „focus group“ auf Google Scholar durchgeführt. Gefunden Treffer, welche für die vorliegende Arbeit als relevant eingestuft wurden, werden in Folge kurz beschrieben. Als relevant wurden Treffer

eingestuft, bei denen eine Fokusgruppe, zur Erhebung von qualitativen Informationen im Kontext von intelligenten Systemen in Smart Homes eingesetzt wurden.

Sixsmith und Johnson (2004) setzten in ihrer Arbeit eine Fokusgruppe ein, um den qualitativen Nutzen und das Potential der Lösung von den Nutzern zu erheben. Sie haben ihr System dabei, wie in der vorliegenden Arbeit, in einer Installation evaluiert und die Fokusgruppe als Ergänzung verwendet.

Demiris et al. (2006) beschreiben in Ihrer Arbeit ein Framework für die Implementierung eines Sensoren Netzwerkes zur Überwachung von älteren Personen. Dabei nutzten sie Beobachtungen in den Wohneinheiten der älteren Personen und Fokusgruppen zur Erhebung der Daten.

Gill, Yang, Yao und Lu (2009) entwarfen ein System zur Automatisierung im Smart Home. Sie nutzten Fokusgruppen, um qualitatives Feedback zu ihrem Systementwurf von den Anwendern und Anwenderinnen einzuholen und so ihre Evaluierung durch qualitative Daten zu ergänzen.

Zusammengefasst finden sich in der Literatur eine große Zahl an Publikationen, welche im Kontext von Smart Homes, Fokusgruppen für die qualitative Datenerhebung verwenden. Viele dieser Arbeiten verwendeten Fokusgruppen als Methode, um Anforderungen, zu erheben oder allgemeine Meinungen zu Themen rund um das Thema Smart Home, zu evaluieren. Es gibt aber auch Arbeiten in denen, wie bei den drei zuvor angeführten Arbeiten, Fokusgruppen als ergänzende Methode zur Erhebung von qualitativen Daten eingesetzt haben. In diesen Fällen sind Fokusgruppen parallel oder im Anschluss an andere Evaluierungen eingesetzt worden.

Für die vorliegende Arbeit wird eine Fokusgruppe mit den Teilnehmern der Fallstudie, im Anschluss an die beiden Iterationen der Evaluierungen in der Fallstudie, durchgeführt. Aufgrund der Limitierung durch den einen vorhandenen Fall, wird auch die Fokusgruppe als einfache Fokusgruppe aufgebaut.

Angelehnt an vorgeschlagene Fragenformulierungen von Krueger und Casey (2014) wurden die folgenden einleitenden Fragen für die Diskussion definiert. Die Fragen wurden dabei bewusst sehr offen gestaltet um die Diskussion nicht einzuschränken.

- Was ist ihre Erfahrung mit dem in der Fallstudie evaluierten Prototyp?
- Wie würden sie die Vorhersagequalität der beiden Iterationen einschätzen?
- Sehen sie einen Nutzen an einem solchen Vorhersage- und Entscheidungssystem?
- Waren die Fehler in den Vorhersagen für eine tägliche Anwendung annehmbar?
- Was hat sie während der Evaluierung am meisten gestört?

Als Moderator für die Fokusgruppe wurde eine externe Person mit Fachwissen zu Data Mining und Smart Homes engagiert, da der Autor der vorliegenden Arbeit selbst an der Fokusgruppe teilgenommen hat. Als Modus für die Fokusgruppe wurde eine Präsenzdiskussion im ruhigen Umfeld gewählt.

5.4 Durchführung der Fallstudie und des Interviews

Dieser Abschnitt dokumentiert den Ablauf und die Durchführung der, zuvor beschriebenen, Evaluierung des Prototyps der vorliegenden Arbeit.

Zur Datensammlung wurden ab 01.07.2017, bis zum Ende der Fallstudie am 19.11.2017, alle Ereignisse im Smart Home, welche relevant für die, in Abschnitt 4.2.2, genannten Attribute waren, aufgezeichnet. Diese Datensammlung, welche in den Datenbanksystemen des Smart Home durchgeführt wurde, bildete die Basis für die Erzeugung der Trainingsdatensätze in der Evaluierung der vorliegenden Arbeit.

Für den Aufbau der Data Mining Modelle für die erste Iteration der Fallstudie wurde aus diesen Daten ein Trainingsdatensatz erstellt, welcher den Zeitraum 01.07.2017 bis 15.09.2017 enthält. Ursprünglich war geplant, den Zeitraum bis 31.08.2018 zu anzusetzen. Aufgrund einer Abwesenheit, der Bewohner des Smart Homes, Ende August und der damit verbundenen Abweichung in der typischen Beleuchtungsnutzung, wurde der Zeitraum um zwei Wochen verlängert. Der daraus generierte Trainingsdatensatz umfasst 87.840 Instanzen. Die Verteilung der Klassen „Beleuchtung Ein“ und „Beleuchtung Aus“, ist in diesem Datensatz sehr ungleich verteilt. Im Durschnitt über alle sechs vorherzusagenden Beleuchtungen sind 95,7% der Instanzen mit der Klasse „Beleuchtung Aus“ und 4,3% der Instanzen mit der Klasse „Beleuchtung Ein“ vorhanden. Dieser Datensatz wurde verwendet, um die Data Mining Modelle für die erste Iteration der Fallstudie zu erstellen. Die Evaluierung mit einer zehnfachen Kreuzvalidierung hat Vorhersagewerte, wie im letzten Durchlauf in Abschnitt 4.2.4, ergeben. Die Detailergebnisse dieser Kreuzvalidierung sind im Anhang, in Tabelle 16, zu finden. Die sechs erstellten Data Mining Modelle wurden im Prototypen installiert und in Betrieb genommen.

Vor der Durchführung der ersten Iteration, wurden die Bewohner und Bewohnerinnen des Smart Home über den Ablauf instruiert. Sie wurden angewiesen, Anomalien und Fehler in der Vorhersage des Prototyps in einem Protokoll aufzuzeichnen, sowie ein Tagebuch über die empfundene Vorhersagequalität des Prototyps zu führen. Die Bewohner und Bewohnerinnen wurden instruiert, als Anomalien sowohl Ereignisse aufzuzeichnen, bei denen der Prototyp den Status einer Beleuchtung unerwünscht verändert hat, als auch Ereignisse, bei denen der Status einer Beleuchtung nicht wie gewünscht geändert wurde. Zu einem Eintrag im Protokoll müssen folgende Daten aufgezeichnet werden: Zeitpunkt, Raum, Beleuchtung und Beschreibung der Anomalie. Für die Führung des Tagebuchs wurden keine Formvorgaben gemacht.

Der erste Durchlauf wurde im Zeitraum 06.11.2017 bis 12.11.2017 durchgeführt. In diesem Zeitraum waren keine geplanten Abwesenheiten oder ähnliches der Bewohner und Bewohnerinnen geplant. Aus den dabei gesammelten Daten und dem begleitenden Protokoll über Anomalien, wurde ein Testdatensatz zur Auswertung der Vorhersagequalität erstellt. Im Zuge dieses Durchlaufes wurden 283 Vorhersagen für Beleuchtungen beobachtet. Die Auswertung der Daten dieses Durchlaufens und die Interpretation der Ergebnisse sind im nachfolgenden Kapitel beschrieben.

Direkt anschließend, an diesen ersten Durchlauf, wurde ein zweiter Durchlauf durchgeführt. Bei diesem Durchlauf wurde die Vorhersagequalität von, mit aktuellen Daten, aktualisierten Data

Mining Modellen evaluiert. Als Vorbereitung für diesen zweiten Durchlauf wurden die Data Mining Modelle aktualisiert. Dabei wurde ein Trainingsdatensatz generiert, welcher den Zeitraum von 01.07.2017 bis 12.11.2017 beinhaltet. Der, für die zweite Iteration generierte Trainingsdatensatz, beinhaltet 155.520 Instanzen. Wie bei der ersten Iteration sind die beiden vorherzusagenden Klassen ungleich verteilt. In diesem Trainingsdatensatz sind 7% der enthaltenen Instanzen mit „Beleuchtung Ein“ und 93% der Instanzen mit „Beleuchtung Aus“ klassifiziert. Durch die Erhöhung der Instanzen mit der Klasse „Beleuchtung Ein“ ist erkennbar, dass die Beleuchtung in diesem Zeitraum stärker genutzt wird, als in der ersten Iteration. Die neu erstellten und damit aktualisierten Data Mining Modelle wurden im Prototyp installiert und in Betrieb genommen. Die Detailergebnisse dieser Kreuzvalidierung sind im Anhang, in Tabelle 17, zu finden. Für die Durchführung dieser Iteration galten dieselben Instruktionen und Vorgaben, wie für die erste Iteration. Im Zuge dieses Durchlaufes wurden 331 Vorhersagen für Beleuchtungen beobachtet. Die Auswertung der Daten dieses Durchlaufens und die Interpretation der Ergebnisse sind im nachfolgenden Kapitel beschrieben.

Um zu überprüfen, ob die, laut Abschnitt 5.1.2, erwartete Änderung im Benützungsverhalten der Beleuchtungen auch eingetreten ist, wurden die durchschnittlichen Beleuchtungsaktivierungen ausgewertet. Es wurden die durchschnittlichen Beleuchtungsaktivierungen pro Tag in den Zeiträumen von 15.07.2017 bis 31.07.2017, sowie von 30.10.2017 bis 05.11.2017, ausgewertet und verglichen. Bei dieser Auswertung wurde eine klar häufigere Nutzung der Beleuchtungen im November, im Vergleich zum untersuchten Zeitraum im Juli festgestellt. Die Anzahl der Beleuchtungsaktivierungen hat sich im Mittelwert über alle sechs beobachteten Beleuchtungen zwischen den beiden Zeiträumen um 76% erhöht. Bei allen sechs Beleuchtungen wurden Steigerungen festgestellt, besonders auffallend war die Beleuchtung im Badezimmer wo sich die Anzahl der täglichen Beleuchtungsaktivierungen fast verdoppelt hat. Durch die Auswertung dieser Steigerung konnte bestätigt werden, dass der erwartete Effekt, welcher als Grundlage für die Evaluierung der Lernfähigkeit definiert wurde, auch eingetreten ist. Die Tabelle mit der Gegenüberstellung der durchschnittlichen Beleuchtungsaktivierungen pro Tag in beiden Zeiträumen ist im Anhang, in Tabelle 18 zu finden.

Im Anschluss an die Durchführung der Fallstudie wurde eine Fokusgruppe mit den Teilnehmern und Teilnehmerinnen abgehalten, um die qualitativen Informationen aus dem begleitenden Tagebuch durch eine Diskussion zu ergänzen. Die Diskussion wurde von einem externen Moderator mit den definierten Einleitungsfragen eingeleitet, die Diskussion drehte sich primär um diese einleitenden Fragen. Aufgrund der sehr kleinen Gruppe mit zwei Personen war es nicht möglich, eine weiterführende Diskussion anzustoßen.

5.5 Zusammenfassung

In diesem Kapitel wurde die Evaluierung von Data Mining Modellen und deren Lernfähigkeit im Bereich des Smart Home, durch eine Literaturrecherche erörtert. Auf Basis der dabei gewonnen Erkenntnisse wurde anschließend eine einfache Fallstudie zur Evaluierung des Prototyps der vorliegenden Arbeit definiert und geplant. Ergänzend zur Fallstudie wurde eine Fokusgruppe

zur Erhebung von qualitativen Daten zur Evaluierung definiert. Im Abschluss des Kapitels wurden die Fallstudie, sowie die Fokusgruppe durchgeführt und die Ergebnisse dokumentiert.

Für die Evaluierung von Data Mining Modellen, im Bereich des Smart Homes, ist es notwendig zuerst ein initiales Datenmodell, auf Basis von über einen definierten Zeitraum gesammelten Daten, zu trainieren. Dieses Data Mining Modell wird dann in das Smart Home integriert und im Feld erprobt. Dabei werden Daten zu den getroffenen Vorhersagen gesammelt. Anschließend werden diese Daten mit Beobachtungen um die Informationen zu manuellen Korrekturen der Entscheidungen des Data Mining Modells ergänzt. Auf Basis dieser Gegenüberstellung und der Auswertung der Vorhersagewerte des Data Mining Modells ist ein Schluss zu Vorhersageleistung des Data Mining Modells im jeweiligen Anwendungsfall möglich.

Für die Evaluierung der Lernfähigkeit von Data Mining Modellen, ist es notwendig, eine bestimmte Veränderung im Smart Home zu beobachten. Im Fall der vorliegenden Arbeit wird der natürliche Verlauf der täglichen Tageslichtzeit als Veränderung genutzt. Dazu müssen zwei Evaluierungen durchgeführt werden. Im Vorfeld werden Daten für das initiale Training des Data Mining Modells gesammelt. Nach einem angemessenen Zeitraum hat sich eine entsprechende Benützungsveränderung der Beleuchtungen durch die geänderte Tageslichtzeit eingestellt. Wenn das passiert ist, wird das initiale Data Mining Modell mit den neuen Testdaten evaluiert. Auf diesem Weg kann die Vorhersageleistung des initialen Modells mit den neuen Daten festgestellt werden. Anschließend wird ein neues Data Mining Modell trainiert, welches kontinuierlich alle Daten im Zeitraum zwischen initialem Training und Evaluierung enthält und ebenso trainiert. Mit diesem Data Mining Modell kann die Vorhersageleistung eines kontinuierlich trainierten Data Mining Modells ermittelt werden. Der Vergleich der Evaluierungen lässt Schlüsse auf die Lernfähigkeit zu.

Für die Durchführung der Evaluierung wurde eine einfache Fallstudie nach Yin (2013) definiert. Für das Vorgehen zur Umsetzung wurde ein Prozess, angelehnt an den Prozess von Noor, (2008) definiert. Als Fall für die Fallstudie wurde das Smart Home des Autors der vorliegenden Arbeit gewählt. Die Fallstudie baut auf der in Abschnitt 1.2 aufgestellte Forschungsfrage und dazugehörenden Hypothesen auf. Der Fall wird in zwei Iteration zu je einer Woche durchlaufen, um eine Evaluierung der Lernfähigkeit, wie zuvor beschrieben, zu ermöglichen. Während dieser Evaluierungsperiode werden die notwendigen Daten für die Auswertung im Smart Home und durch die Bewohner und Bewohnerinnen erhoben. Die Fallstudie wird mit einem Tagebuch begleitet, in dem qualitative Eindrücke und Meinungen zum Prototyp und der Vorhersageleistung erhoben werden.

Im Anschluss an die Durchführung der Fallstudie wurde eine Fokusgruppe definiert, um die erhobenen Informationen zu ergänzen. Der Fokus dabei liegt dabei auf dem qualitativen Eindruck der Vorhersageleistung, sowie dem Nutzen eines solchen Vorhersage- und Entscheidungssystems.

Im Abschluss wurde die Durchführung der Fallstudie, sowie der anschließenden Fokusgruppe dokumentiert. In der ersten Iteration der Fallstudie wurden 283 und in der zweiten Iteration 331 Vorhersagen beobachtet welche ausgewertet werden können. Die Fokusgruppe drehte sich primär um die Einleitungsfragen, eine weiter reichende Diskussion ist nicht entstanden.

6 ERGEBNISSE

In diesem Kapitel werden die in der Fallstudie, in dem begleitenden Tagebuch, sowie die in der anschließenden Fokusgruppe erhobenen Daten aufbereitet, ausgewertet und die Ergebnisse dokumentiert. Die Auswertung der erhobenen Daten in der Fallstudie bildet die Grundlage zur Beantwortung der Forschungsfrage und wird von den erhobenen qualitativen Daten aus dem Tagebuch und der Fokusgruppe ergänzt. Abschließend werden die Ergebnisse zusammengefasst und die Forschungsfrage beantwortet.

6.1 Auswertung der Fallstudie

Die, während der beiden Iterationen der ersten Fallstudie gesammelten Daten wurden je Iteration ausgewertet. Anschließend wurden die Ergebnisse der beiden Iterationen zur Beantwortung der Hypothesen und der Forschungsfrage verglichen.

Konkret wurde ein Protokoll über sämtliche, im Beobachtungszeitraum durchgeführten Vorhersagen des Prototyps erstellt und mit den tatsächlichen Veränderungen im Zustand der Beleuchtungen, im Smart Home, kombiniert. Dieses Protokoll wurde anschließend mit den Informationen aus dem begleitenden Protokoll über Anomalien vervollständigt. Auf Basis dieser Daten wurde ein Testdatensatz erstellt, welcher je durchgeführter Vorhersage, eine Instanz enthält, welche mit der korrekten Klassifizierung versehen wurde. Diese Klassifizierung erfolgte auf der, in Abschnitt 5.2.2, beschriebenen Logik. So wurden zuerst alle Instanzen automatisch auf Basis der Daten klassifiziert und diese Klassifizierung anschließend mit Hilfe des begleitenden Protokolls korrigiert. Anhand dieser Testdatensätze wurde anschließend eine Evaluierung mit den Data Mining Modellen durchgeführt, um die tatsächlichen Vorhersagewerte zu ermitteln.

Die Vorhersagewerte für die erste Iteration der Fallstudie sind in Tabelle 14 dargestellt. Die Vorhersagewerte für die Klasse „Beleuchtung Aus“ (OFF) sind sehr gut ausgefallen. Hier konnten alle sechs Data Mining Modelle eine gute Rate von korrekten Vorhersagen (True Positiv) erreichen. Nur das Modell 4, für die Vorhersage der Beleuchtung beim Fernseher, hat eine Rate für die korrekte Vorhersage dieser Klasse mit unter 95% erreicht. Für die Klasse „Beleuchtung Ein“ (ON) konnten im ersten Durchlauf keine guten Vorhersagewerte erreicht werden. Mit Ausnahme von Data Mining Modell 4, für die Vorhersage der Beleuchtung beim Fernseher, konnte kein Data Mining Modell eine höhere Rate für korrekte Vorhersagen als 17% erreichen. Dieses Ergebnis wurde erwartet, da sich die Beleuchtungsnutzung im Zeitraum zwischen dem Datensatz mit dem das Modell trainiert wurde und dem Testdatensatz, aus dem Evaluierungszeitraum, verändert hat. Die, trotz dieser Veränderung, sehr guten Vorhersagewerte für die Klasse „Beleuchtung Aus“ lassen sich damit erklären, dass die

Beleuchtung im Aufzeichnungszeitraum für den Trainingsdatensatz, weniger im Einsatz war, als im Evaluierungszeitraum. Das bedeutet, dass zu einem Zeitpunkt an, dem zum Evaluierungszeitraum eine Beleuchtung inaktiv war, die Wahrscheinlichkeit hoch ist, dass die Beleuchtung zu einem ähnlichen Zeitpunkt im Trainingszeitraum ebenso inaktiv war. Wohingegen bei der Klasse „Beleuchtung Ein“, durch die verstärkte Nutzung der Beleuchtungen, eine größere Veränderung einhergeht und daher eine große Abweichung entsteht. Das umgekehrte Verhalten, dass sich die Vorhersagewerte für die Klasse „Beleuchtung Aus“ stark verschlechtern, aber die Vorhersagewerte für die Klasse „Beleuchtung Ein“ relativ stabil bleiben, könnte im Frühjahr, wenn die Tage wieder länger werden und die Nutzung der Beleuchtungen wieder zurückgehen, festgestellt werden. Diese Annahme wird in der vorliegenden Arbeit nicht weiter behandelt.

Modell	TP „OFF“	FP „OFF“	Präzision „OFF“	Trefferquote „OFF“	TP „ON“	FP „ON“	Präzision „ON“	Trefferquote „ON“
1	100%	97,6%	50,6%	100%	2,4%	0%	100%	2,4%
2	100%	82,8%	54,7%	100%	17,2%	0%	100%	17,2%
3	95,2%	85,7%	52,6%	95,2%	14,3%	4,8%	75,0%	14,3%
4	93,3%	46,7%	66,7%	93,3%	53,3%	6,7%	53,3%	50,9%
5	100%	0,0%	50,0%	100%	0,0%	0,0%	100,0%	0,0%
6	100%	94,4%	100%	100%	5,6%	0,0%	100,0%	10,5%

Tabelle 14: Vorhersagewerte der ersten Iteration der Fallstudie

In Summe kann zur ersten Iteration der Fallstudie gesagt werden, dass die Data Mining Modelle im Prototypen sehr gut beim Deaktivieren von Beleuchtungen abgeschnitten haben, dass sie aber schlecht in der korrekten Vorhersage einer Beleuchtungsaktivierung abgeschnitten haben.

In Tabelle 15 sind die Vorhersagewerte der zweiten Iteration der Fallstudie dargestellt. Die Vorhersagewerte für die Klasse „Beleuchtung Ein“, konnten gegenüber der ersten Iteration der Fallstudie, gesteigert werden. Die Vorhersagewerte der Klasse „Beleuchtung Aus“ hingegen, haben sich teilweise verschlechtert. Die Vorhersagewerte der Klasse „Beleuchtung Ein“, haben sich für alle sechs Data Mining Modelle verbessert. Die größten Steigerungen konnte für die Data Mining Modelle 1 und 6 erreicht werden. Die Rate für korrekte Vorhersagen für die Klasse „Beleuchtung Ein“, für das Data Mining Modell 1, zur Vorhersage der Beleuchtung im Badezimmer, konnte um 29,3% gesteigert werden. Die Rate für korrekte Vorhersagen für die Klasse „Beleuchtung Ein“, für das Data Mining Modell 6, zur Vorhersage der Beleuchtung in der Spielecke, konnte um 33,3% gesteigert werden. Wie in Tabelle 15 ersichtlich, wurde für die Beleuchtung im Bad eine große Änderung im Nutzungsverhalten im Zeitraum festgestellt. Auch für die Beleuchtung beim Fernseher, welches von Data Mining Modell 4 vorhergesagt wird, wurde eine große Veränderung im Nutzungsverhalten identifiziert. Das Benützungsverhalten dieser Beleuchtung konnte aber in beiden Iterationen ähnlich gut vorhergesagt werden. Die Rate für korrekte Vorhersagen der Klasse „Beleuchtung Ein“ konnte bei diesem Modell nur um 3,3% gesteigert werden. Die Vorhersagewerte für die Klasse „Beleuchtung Aus“, haben sich bei

den Data Mining Modellen 1, 2, 3 und 4 leicht verschlechtert. Die Verbesserung der Werte für die Klasse „Beleuchtung Ein“ überwiegt jedoch deutlich.

Modell	TP „OFF“	FP „OFF“	Präzision „OFF“	Trefferquote „OFF“	TP „ON“	FP „ON“	Präzision „ON“	Trefferquote „ON“
1	95,9%	68,6%	54,8%	82,9%	31,7%	4,1%	88.5%	31,7%
2	93,3%	58,6%	57,5%	79,3%	41,4%	6,7%	92,2%	41,4%
3	95,5%	66,7%	57,6%	90,5%	33,3%	4,5%	95,8%	33,3%
4	92,2%	43,3%	67,5%	90,0%	56,7%	3,0%	95,0%	56,7%
5	100%	75,0%	57,1%	100%	25,0%	0,0%	100%	25,0%
6	100%	61,1%	62,1%	100%	38,9%	0,0%	100%	38,9%

Tabelle 15: Vorhersagewerte der zweiten Iteration der Fallstudie

Zusammengefasst, wurden bei der zweiten Iteration der Fallstudie durchgehend bessere Vorhersagewerte für die Klasse „Beleuchtung Ein“ festgestellt. Für die Klasse „Beleuchtung Aus“ sind die Vorhersagewerte bei vier von sechs Data Mining Modellen stabil geblieben und haben sich bei zwei Modellen verschlechtert.

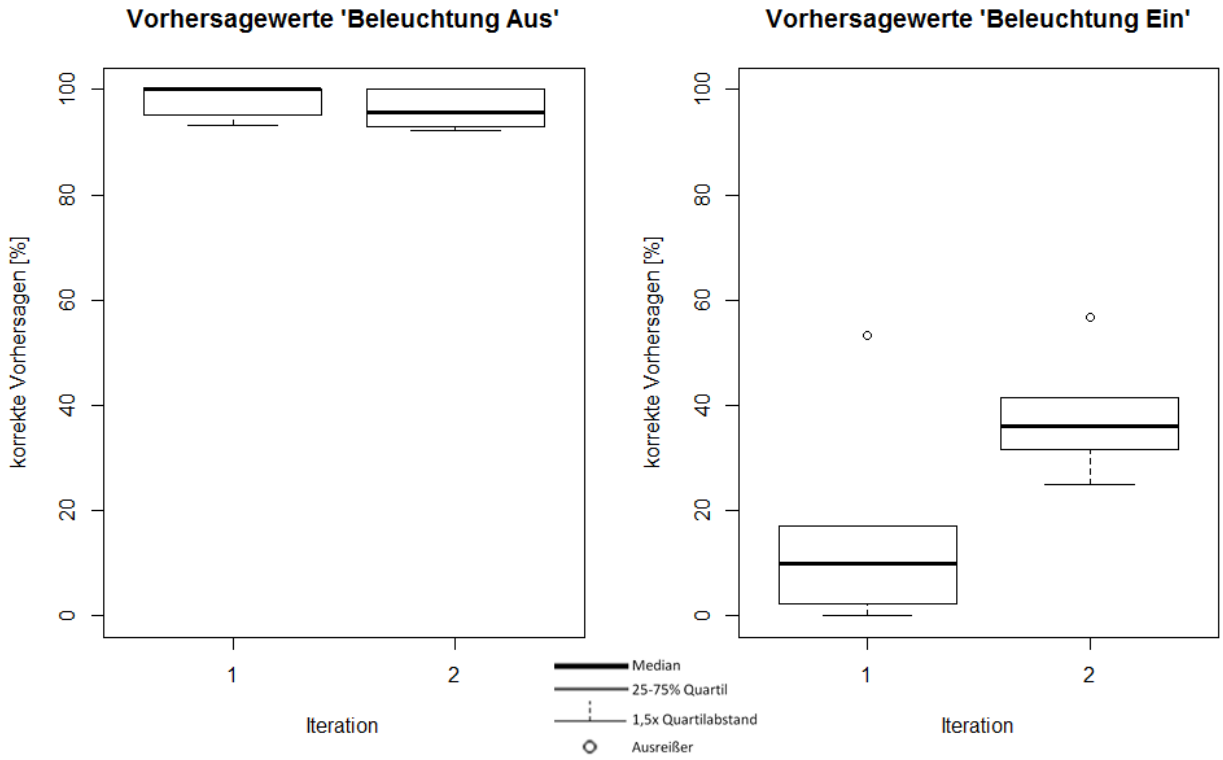

Abbildung 8: Gesamtvorhersagewerte der beiden Iterationen

Bei einem Vergleich der Gesamtvorhersageleistung der beiden Iterationen konnte festgestellt werden, dass sich die Vorhersagewerte für die Klasse „Beleuchtung Ein“ im Mittel über alle sechs Data Mining Modelle um 22,37% verbessert haben. Die Vorhersagewerte für die Klasse „Beleuchtung Aus“, haben sich im Mittel über alle sechs Data Mining Modelle um 1,93%

verschlechtert. In Summe konnte eine Verbesserung der Vorhersagewerte in der zweiten Iteration festgestellt werden. In Abbildung 8 sind Gesamtvorhersagewerte je Klasse und Iteration dargestellt. In dieser Darstellung ist ersichtlich, wie sich die Vorhersageleistung der Klasse „Beleuchtung Ein“ in der zweiten Iteration verbessert hat. Auch die leichte Verschlechterung der Vorhersagewerte für die Klasse „Beleuchtung Aus“, in der zweiten Iteration, kann in dieser Abbildung erkannt werden.

Die, in den Iterationen beobachteten, Vorhersagewerte für die Klasse „Beleuchtung Ein“, sind wesentlich schlechter ausgefallen, als die Vorhersagewerte der jeweiligen Kreuzevaluierung mit dem Trainingsdatensatz. Dieser Unterschied lässt sich damit erklären, dass das initiale Training mit einem Datensatz durchgeführt wurde, welcher eine Trainingsinstanz pro Minute beinhaltet hat. Die Evaluierung in der Fallstudie wurde jedoch mit konkreten Ereignissen durchgeführt. Die unterschiedlich guten Vorhersagewerte lassen sich durch diesen Unterschied erklären.

6.2 Auswertung des begleitenden Tagebuchs und der Fokusgruppe

Das begleitende Tagebuchprotokoll beider Iterationen wurde ausgewertet, um einen qualitativen Eindruck zur Vorhersagequalität, durch die Bewohner und Bewohnerinnen des Smart Homes, zu erhalten. Die beiden Bewohner haben zusammen, ein Tagebuch über ihren Eindruck zur Vorhersageleistung am jeweiligen Tag geführt. Eine Abschrift des handschriftlichen Tagebuches für beide Iterationen ist im Anhang abgelegt.

Die Aufzeichnungen der ersten Iteration, untermauern die Ergebnisse der Auswertung aus der Fallstudie, die Aktivierung der Beleuchtung, also die Vorhersage der Klasse „Beleuchtung Ein“, hat nach dem Empfinden der Bewohner und Bewohnerinnen so gut wie gar nicht funktioniert. Es ist durchaus aufgefallen, dass die Deaktivierung der Beleuchtung, also die korrekte Vorhersage der Klasse „Beleuchtung Aus“, funktioniert hat, jedoch wurde diese Leistung sehr stark durch die äußerst schlechte Vorhersagequalität der anderen Klasse zunichte gemacht. Die Tagebucheinträge vermitteln das Gefühl, als ob durch die falschen Vorhersagen der Klasse „Beleuchtung Ein“, die gesamte Leistung des Prototyps, als schlecht wahrgenommen wird. Besonders negativ ist aufgefallen, dass die Bewohner, durch falsche Deaktivierungen von Beleuchtungen, plötzlich in dunklen Räumen gestanden sind.

Auch in den Tagebucheinträgen der zweiten Iteration, spiegeln sich die, in der Fallstudie erhobenen Vorhersagewerte wieder. Es wurde wahrgenommen, dass sich die Vorhersageleistung der Klasse „Beleuchtung Ein“ verbessert hat. Bei ein paar Tagebucheinträgen konnte auch der Eindruck gewonnen werden, dass die Vorhersageleistung von den Bewohnern akzeptiert und der Prototyp angenommen wurde. Nicht funktionierende Vorhersagen zur Klasse „Beleuchtung Ein“, ziehen sich durch das gesamte Tagebuch. Offensichtlich wurde Verbesserung gegenüber der ersten Iteration festgestellt, die Vorhersagequalität ist in Summe aber nicht ausreichend hoch, damit der Prototyp von den Bewohnern erfolgreich angenommen wird.

Bei der Fokusgruppe wurde primär auf die, in Abschnitt 5.3, definierten Fragen eingegangen. So wurde festgestellt, dass die Erfahrung mit dem Prototyp während der Fallstudie,

durchgängig sehr negativ war. Primär wurden die schlechten Vorhersagen der Klasse „Beleuchtung Ein“ kritisiert. Die Fehler wurden als nicht annehmbar eingestuft. Für die Bewohner und Bewohnerinnen ist die Vorhersageleistung der Klasse „Beleuchtung Ein“, als wesentlich wichtiger, als die Vorhersageleistung der Klasse „Beleuchtung Aus“, eingestuft worden. Es ist akzeptabel, wenn das Licht bei Verlassen des Raumes eingeschalten bleibt. Es ist aber nicht akzeptabel, wenn das Licht bei Betreten des Raumes ausgeschalten wird, obwohl man Licht benötigt. In Summe wurde die Idee des Prototyps als gut empfunden. Die Vorhersageleistung wurde während beider Iterationen, als zu schlecht für den praktischen Einsatz eingestuft.

Zusammengefasst haben die qualitativen Datenerhebungen, die quantitativen Ergebnisse aus subjektiver Sicht der Bewohner und Bewohnerinnen bestätigt. In der zweiten Iteration, wurde eine bessere Vorhersageleistung, als in der ersten Iteration wahrgenommen. In Summe war die Fehlerrate in beiden Iterationen aber viel zu hoch. So konnte der Prototyp von den Bewohnern, nicht als praktikabel und hilfreich angenommen werden. Die falschen Vorhersagen standen während der gesamten Evaluierung im Vordergrund und haben die Wahrnehmung sehr stark negativ beeinflusst.

6.3 Zusammenfassung

In diesem Kapitel wurde zuerst die quantitative Datenerhebung der Fallstudie ausgewertet und die Daten zur Beantwortung der Forschungsfrage erhoben. Die im Smart Home gesammelten Daten zu durchgeführten Beleuchtungsaktivierungen und durchgeführten Vorhersageereignisse, wurden je Iteration, mit den manuellen Anomalieprotokollen kombiniert, um einen Evaluierungsdatensatz zu erstellen. Die dabei entstanden Evaluierungsdatensätze wurden verwendet, um die Vorhersagequalität der Data Mining Modelle zu ermitteln. Diese Vorhersageleistungen wurden, je vorherzusagender Klasse und Iteration, ausgewertet und gegenübergestellt.

In der ersten Iteration wurde eine sehr gute Vorhersageleistung für die Klasse „Beleuchtung Aus“ festgestellt. Im Durchschnitt über alle sechs Data Mining Modelle wurden 98% der Vorhersagen für diese Klasse korrekt durchgeführt. Die Vorhersageleistung der Klasse „Beleuchtung Ein“ hat in dieser Iteration nur sehr schlechte Werte erreicht. Im Durchschnitt über alle sechs Data Mining Modelle, wurden nur 15, 5% der Vorhersagen korrekt durchgeführt. In der zweiten Iteration konnte die Vorhersageleistung der Klasse „Beleuchtung Ein“ auf 37,83% korrekte Vorhersagen, gesteigert werden. Die Vorhersageleistung der Klasse „Beleuchtung Aus“ hingegen, ist leicht auf 96,2% abgesunken.

Zusammengefasst wurde festgestellt, dass durch den Einsatz von aktualisierten Data Mining Modellen, in der zweiten Iteration, die Vorhersageleistung des Prototyps gesteigert werden konnte. Konkret konnte die Vorhersageleistung der Klasse „Beleuchtung Ein“ zwischen den beiden Iterationen um 22,37% gesteigert werden.

Die ergänzenden qualitativen Datenerhebungen, durch das Tagebuch und die abschließende Fokusgruppe, haben diese Ergebnisse in der Wahrnehmung der Bewohner bestätigt. Hier

wurde jedoch auch festgestellt, dass die Vorhersagequalität für einen praktischen Einsatz viel zu niedrig war. Damit ein Vorhersage- und Entscheidungssystem für diesen Anwendungsfall von den Benutzern angenommen wird, muss die Vorhersagqualität für die Klasse „Beleuchtung Ein“ deutlich gesteigert werden.

7 ZUSAMMENFASSUNG

In diesem Kapitel werden die Ergebnisse und Erkenntnisse der vorliegenden Kapitel zusammengefasst. Im ersten Teil wird die gewählte Vorgehensweise der vorliegenden Arbeit reflektiert. Anschließend werden die Ergebnisse der Fallstudie, sowie der begleitenden Methoden als Antwort auf die Forschungsfrage zusammengefasst und dokumentiert. Zum Abschluss werden der Nutzen der vorliegenden Arbeit und ein Ausblick auf mögliche weiterführende Forschungsziele, in diesem Bereich, beschrieben.

7.1 Reflexion der Vorgehensweise

Im ersten Kapitel wurden die Begriffe Data Mining, Machine Learning und Smart Home für die vorliegende Arbeit definiert. Der Begriff Data Mining wurde als Prozess zur Wissensgenerierung aus großen Datenmengen und der Begriff Machine Learning als Anwendung eines konkreten Algorithmus in diesem Wissensgenerierungsprozess, definiert. Unter dem Begriff Smart Home wurden, in der vorliegenden Arbeit Wohneinheiten eingeordnet, welche mit vernetzten Sensoren und Aktoren zur Steuerung und Automatisierung ausgerüstet sind. Diese Wohneinheiten werden durch intelligente Assistenzsysteme unterstützt, welche den Wohnkomfort durch intelligente Automatisierung von Aufgaben im Smart Home erhöhen.

Für die Erörterung der Vorgehensweise zur praktischen Anwendung von Data Mining wurde auf den weit verbreiteten und standardisierten Data Mining Prozess CRISP-DM zurückgegriffen. Dieser Prozess beschreibt ein iteratives Vorgehen zur Lösung von Data Mining Problemstellungen. Das Vorgehen dieses Prozesses wurde durch eine Literaturrecherche geklärt und mit anderen Vorgehensmodellen, für die Anwendung von Data Mining verglichen. Das iterative Vorgehen des CRISP-DM Prozesses wurde für die Umsetzung des Prototyps in der vorliegenden Arbeit, übernommen und angewandt.

Anschließend wurden mit einer Literaturrecherche Data Mining Verfahren ermittelt, welche für die Lösung von Klassifizierungsproblemen im Bereich des Smart Homes geeignet sind. Dabei wurde der Fokus auf Klassifizierungsverfahren gelegt, welche Zwei-Klassen Probleme lösen können. Diese Literaturrecherche stützte sich auf Fachbücher und Standardwerke zum Thema Data Mining und auf Publikationen, welche sich mit der Verbreitung und Popularität von Data Mining Verfahren beschäftigen. Durch die Ergebnisse dieser Literaturrecherche wurden die sieben Data Mining Verfahren, „Entscheidungsbäume“, „Stützvektormaschinen“, „Nächste Nachbarn Klassifikation“, „Naive Bayes Klassifikation“, „neuronale Netzwerke“, „Regelinduktion“ und „lineare Modelle“, als für den Anwendungsfall, der vorliegenden Arbeit relevant eingestuft. Die Funktionsweise dieser Data Mining Verfahren, wurde im Anschluss durch eine Literaturrecherche erhoben und eine Klassifizierung dieser Data Mining Verfahren, wurde

durchgeführt, um eine Basis für eine Einschränkung und Auswahl zu schaffen. Auf Basis dieser, wurden die Data Mining Verfahren anschließend eingeschränkt. Die beiden Data Mining Verfahren „naive Bayes Klassifizierung“ und „nächster Nachbarn Klassifikation“, wurden aufgrund der Klassifikation für die vorliegende Arbeit ausgeschlossen, weil sie die Anforderungen aus dem Anwendungsfall der vorliegenden Arbeit nicht voll unterstützen. Für die verbleibenden fünf Data Mining Verfahren wurde ein praktischer Vergleich durchgeführt, um zu ermitteln, welches Data Mining Verfahren für den Anwendungsfall der vorliegenden Arbeit am besten geeignet ist. Zu diesem Zweck wurde für jedes dieser Data Mining Verfahren, ein Data Mining Modell mit Daten aus dem Anwendungsfall trainiert und evaluiert. Das Data Mining Verfahren Entscheidungsbäume, hat bei dieser Evaluierung, die besten Vorhersagewerte geliefert und wurde aus diesem Grund für die Anwendung in der vorliegenden Arbeit ausgewählt.

Basierend auf den Technologien, welche im Smart Home des Autors der vorliegenden Arbeit im Einsatz sind, wurden Technologien für die Umsetzung des Prototyps der vorliegenden Arbeit, ausgewählt. Der Prototyp wurde nahtlos in die Infrastruktur des Smart Homes integriert. Dazu wurde die Software Weka als Data Mining Bibliothek und Spring Boot als Basisframework, für die Umsetzung des Prototyps gewählt. Für die Kommunikation mit den anderen beteiligten Smart Home Komponenten, wurde auf den bestehenden MQTT Bus zurückgegriffen. Für die Präsenzerkennung im Smart Home, wurde eine Android App als Basis gewählt, welche mit Hilfe von Eddystone Bluetooth Beacons, die Präsenz des Smart Homes in einem Raum erkennen kann.

Unter Anwendung des CRISP-DM Prozesses wurde mit den gewählten Technologien, ein Prototyp für den vorliegenden Anwendungsfall entwickelt und in das bestehende Smart Home integriert. Im ersten Schritt wurden die fachlichen Anforderungen erhoben, welche das Benützungsverhalten von Beleuchtungen in einer Wohneinheit, beeinflussen. Auf Basis dieser Anforderungen, welche durch eine Literaturrecherche aus dem Bereich der Gebäudesimulation unterstützt wurde, ist in den Daten des Smart Homes nach Datenquellen und Attributen gesucht worden. In Summe wurden in den Datenquellen des Smart Homes, 30 globale und lokale Parameter identifiziert, welche als, für die Data Mining Modelle der vorliegenden Arbeit, relevant eingestuft wurden. Aus diesen Attributen wurde ein Extraktionsprozess zur Erstellung von Trainings- und Evaluierungsdatensätzen definiert. Dieser Prozess ermöglicht es, vollautomatisiert, Datensätze für die Erstellung und Evaluierung von Data Mining Modellen, aus den Datenquellen des Smart Homes zu generieren. Mit Hilfe dieser Datensätze, wurden anschließend sechs Data Mining Modelle für den Prototyp der vorliegenden Arbeit trainiert und evaluiert. Im Zuge dieses Vorgehens wurde in einem iterativen Vorgehen, die Vorhersageleistung der Data Mining Modelle erhöht, in dem die Attribute, welche für das Training verwendet wurden, iterativ analysiert und adaptiert wurden. Abschließend wurden die erstellten Data Mining Modelle in den Prototyp integriert.

Der Funktionsumfang des Prototyps umfasst zwei Hauptprozesse. Einen Prozess zum Erstellen und Aktualisieren der Data Mining Modelle, durch neue Trainingsdaten und einen Prozess zum Treffen von Vorhersagen, durch Einsatz der trainierten Data Mining Modelle. Der erste Prozess wird für die Vorbereitung des Prototyps zur Evaluierung verwendet und manuell angestoßen.

Der zweite Prozess wird aktiv von der Android App angestoßen, wenn ein Bewohner oder eine Bewohnerin des Smart Homes, einen Raum betritt oder verlässt. Durch dieses Ereignis wird mit Hilfe der Data Mining Modelle, eine Vorhersage zum gewünschten Zustand der jeweiligen Beleuchtung getroffen und gegebenenfalls eine Veränderung des Zustandes der Beleuchtung im Smart Home ausgelöst.

Der erstellte Prototyp wurde in das Smart Home zur Evaluierung und Erhebung der Daten, welche für die Beantwortung der Forschungsfrage notwendig sind, integriert. Für die Evaluierung wurde eine einfache Fallstudie durchgeführt. Bei dieser Fallstudie wurden, in zwei Iterationen, mit einer Durchlaufzeit von je einer Woche, einmal die Vorhersageleistung eines einmal erstellten Data Mining Modells und anschließend die Vorhersageleistung eines aktualisierten Data Mining Modells, evaluiert. Diese Evaluierung hatte das Ziel, die Veränderung im Benützungsverhalten von Beleuchtungen zu beobachten, welche durch die natürliche Verschiebung des Tag-Nacht Zyklus im Jahresverlauf entsteht. Die beiden Iterationen wurden von einem Anomalieprotokoll, zur Vervollständigung der Beobachtungen, sowie von einem Tagebuch, zur ergänzenden Erhebung von qualitativen Daten, begleitet. Als Ergänzung zu diesem Vorgehen, wurden im Anschluss an die Fallstudie, eine Fokusgruppe mit den Teilnehmern und Teilnehmerinnen durchgeführt, um eine qualitative Einschätzung zur Vorhersageleistung des Prototyps zu erhalten. Diese erhobenen Daten aus den Beobachtungen in den Iterationen der Fallstudie, aus dem Anomalieprotokoll, sowie aus dem Tagebuch und der abschließenden Fokusgruppe, wurden abschließend zur Beantwortung der Forschungsfrage kombiniert und die beiden Iterationen verglichen.

Die Durchführung der Fallstudie konnte in der vorliegenden Arbeit nur als einfache Fallstudie geplant und durchgeführt werden. Diese Limitierung war durch die Einschränkung auf die, für die Evaluierung zur Verfügung stehenden Smart Homes vorgegeben und konnte im Rahmen der vorliegenden Arbeit nicht erweitert werden. Durch die Betrachtung eines einzelnen Falles, sind die gewonnen Erkenntnisse sehr stark auf diesen Fall bezogen. Für eine aussagekräftigere Evaluierung auf allgemeiner Ebene wäre es notwendig, die Evaluierung auf eine größere Anzahl von Fällen auszuweiten. Eine Erhöhung der Anzahl der Iterationen könnte die Qualität der Beobachtung, über einen längeren Zeitraum, verbessern.

Im Zuge der Erstellung der Data Mining Modelle, ist das extreme Ungleichgewicht der beiden vorherzusagenden Klassen, in den Trainingsdatensätzen aufgefallen. Solche Ungleichgewichte beeinflussen die Vorhersageleistung der Data Mining Modelle. In der Literatur finden sich verschiedene Verfahren, mit denen diese Ungleichgewichte besser behandelt werden können. In der vorliegenden Arbeit wurde auf die Anwendung eines solchen Verfahrens verzichtet, da einfache Data Mining Verfahren, ohne Adaptionen und Erweiterungen, zum Einsatz kommen sollten. Durch die Anwendung von solchen Verfahren, kann die Vorhersageleistung der einzelnen Klassen besser ausbalanciert werden, was unter Umständen, zu einer besseren Vorhersageleistung des Prototyps führen könnte.

Bei der Analyse der Datenquellen und der vorhandenen Sensoren, im verwendeten Smart Home wurde festgestellt, dass nicht alle Räume über die volle Ausprägung an installierten Sensoren verfügen, wodurch in einigen Fällen, auf globale Sensoren zurückgegriffen wurde.

Eine größere Zahl an Sensoren direkt in den Räumen und ein engeres Netz an Sensoren, hätte eine bessere Datenbasis für das Training der Datenmodelle bilden können. Die Gegebenheiten des vorhandenen Smart Homes, haben den Bau des Prototyps auf die bestehenden Sensoren und Daten limitiert.

Bei der abschließenden Fokusgruppe wurde festgestellt, dass die Präsenzerkennung mit den Bluetooth Beacons nicht in allen Fällen reibungslos funktioniert hat. In einigen Situationen bemerkten die Bewohner, eine Zeitdifferenz von mehreren Sekunden zwischen Betreten oder Verlassen eines Raumes und der erwarteten Aktion, durch die Vorhersage des Prototyps. Diese Verzögerung kann negative Einflüsse auf die qualitative Wahrnehmung der Vorhersageleistung bewirken.

7.2 Reflexion der Ergebnisse

Die Forschungsfrage „Welche Auswirkung hat das kontinuierliche Training des Data Mining Modells, für ein Vorhersagesystem zur intelligenten Automatisierung der Beleuchtungssteuerung in einem Smart Home, auf die Fehlerrate des Data Mining Modells?“ kann durch die, in Abschnitt 6.1, durchgeführte Auswertung der Fallstudie beantwortet werden. Im Zuge der Auswertung der Fallstudie wurden die beiden durchgeführten Iterationen gegenübergestellt und verglichen. Durch die Ergebnisse dieses Vergleiches konnte, für den vorliegenden Anwendungsfall, die Hypothese H1 bestätigt und die Hypothese H0 wiederlegt werden. Durch den Einsatz von kontinuierlich Aktualisierten Data Mining Modellen, konnte die Fehlerrate, im Durchschnitt, um 22,37% gesenkt werden.

Die Forschungsfrage kann somit damit beantwortet werden, dass ein kontinuierliches Training von Data Mining Modellen, für ein Vorhersagesystem zur intelligenten Automatisierung der Beleuchtungssteuerung in einem Smart Home, zur Verbesserung der Fehlerraten, im Vergleich zu einmalig trainierten Data Mining Modellen, beiträgt.

Durch die Limitierung auf eine einfache Fallstudie zur Evaluierung, sind die Ergebnisse zwar für den vorliegenden Anwendungsfall gültig, es ist aber nicht möglich eine Aussage mit allgemeiner Gültigkeit zu treffen.

Eine erneute Durchführung der Fallstudie im selben Rahmen, aber einen größeren Zeitraum kann zur Festigung und Ergänzung der Erkenntnisse beitragen. Durch einen längeren Beobachtungszeitraum, würden externe Störeinflüsse eine geringere Auswirkung auf die Evaluierungsergebnisse einzelner Iterationen haben und diese Iterationen somit stabiler vergleichbar machen.

Über diese Erkenntnisse hinaus wurde, durch die Auswertung der qualitativen Evaluierung, anhand des Tagebuches und der abschließenden Fokusgruppe, erhoben, dass die Vorhersagequalität für ein solches Vorhersagesystem sehr gut sein muss, damit es von den Bewohnern und Bewohnerinnen akzeptiert wird. Falsche Vorhersagen hinterlassen sehr schnell den Eindruckt, dass das Vorhersagesystem völlig versagt und nicht funktioniert. Außerdem wurden die auslösenden Ereignisse beim Betreten und Verlassen des Raumes, als nicht

ausreichend angesehen. Ein entsprechendes Vorhersagesystem sollte in der Lage sein, relevante Vorhersagen auch während dem Aufenthalt von Personen im Raum zu erkennen und auszulösen.

7.3 Conclusio

Durch die vorliegende Arbeit hat sich gezeigt, dass Data Mining Modelle zur Automatisierung im Bereich des Smart Homes, durch kontinuierliche Aktualisierung mit neuen gesammelten Daten, zeitnahe auf Veränderungen im Benützungsverhalten des Smart Homes angepasst werden können. Das kontinuierliche Training wirkt sich positiv auf die Verbesserung und Beibehaltung der Vorhersagequalität, von Data Mining Modellen, in diesem Anwendungsfall aus.

Es hat sich auch gezeigt, dass bereits kleine Veränderungen im Benützungsverhalten, wie sie durch den, sich im Jahresverlauf verschiebenden, Tages-Nacht Zyklus entstehen, negative Auswirkungen auf die Vorhersagequalität von Data Mining Modellen zur Automatisierung im Smart Home haben können. Bei Data Mining Modellen für Automatisierungsaufgaben im Bereich des Smart Homes, kann ein kontinuierliches Aktualisieren des Data Mining Modells, durch neue Daten, die Vorhersagequalität verbessern und den negativen Auswirkungen auf die Vorhersagequalität, welche bereits durch kleine Veränderungen im Benützungsverhalten ausgelöst werden können, entgegenwirken.

7.4 Nutzen und Ausblick

Die in der vorliegenden Arbeit gewonnen Erkenntnisse, können in Zukunft beim Bau von Vorhersage- und Entscheidungssystemen, für die Automatisierung im Smart Home, dazu beitragen, geeignete Data Mining Verfahren auszuwählen und zu entscheiden, ob für den jeweiligen Anwendungsfall ein kontinuierliches Training der Data Mining Modelle Vorteile bietet.

Die vorliegende Arbeit zeigt das Potential von einfachen Data Mining Verfahren für solche Aufgaben. Es wurde gezeigt, dass es auch mit einfachen Mitteln möglich ist, ein selbstlernendes Vorhersage- und Entscheidungssystem zur Automatisierung, im Bereich des Smart Home, zu erstellen. Der Einsatz von Rechenintensiven Anwendungen in der Cloud ist für diesen Einsatzbereich keine zwingende Voraussetzung.

Für die weitere Forschung in diesem Bereich könnte die Möglichkeit zur Verbesserung der allgemeinen Vorhersagequalität der Data Mining Modelle, untersucht werden. Für einen praktischen Einsatz dieser Vorgehensweise, muss die Vorhersagequalität noch gesteigert werden. Diese Optimierungen könnten zum Beispiel, durch das hinzufügen weiterer Sensoren, das Ausweiten des betrachteten Zeitraumes oder Optimierung des eingesetzten Data Mining Modells, erreicht werden. Die in der vorliegenden Arbeit gezeigte Vorgehensweise, zum Bau eines solchen Systems, könnte auch in Hinblick auf eine Vergrößerung, durch die Verbindung mehrerer Smart Homes oder für den Einsatz in einer Smart City, adaptiert und angewandt werden. Weiteres Potential wird darin gesehen, das Data Mining Modell, die Datensammlung,

sowie die Entscheidungsfindung, für jeden Bewohner und jede Bewohnerin des Smart Homes zu personalisieren. Auf diese Weise könnte das Vorhersage- und Entscheidungssystem, besser auf die Vorlieben zur Beleuchtungsnutzung einzelner Personen, im Smart Home, Rücksicht nehmen.

Die in der vorliegenden Arbeit angewandte Vorgehensweise zur Evaluierung der Vorhersagequalität von Data Mining Modellen, könnte in zukünftigen Arbeiten, für die Evaluierung in ähnlichen Anwendungsgebieten angewandt werden. In diesem Fall könnte die Belastbarkeit der erhobenen Daten durch eine mehrfache Fallstudie verbessert werden.

ANHANG A - 1. Anhang

Daten zum ersten Durchlauf der Fallstudie

Modell	TP „OFF“	FP „OFF“	Präzision „OFF“	Trefferquote „OFF“	TP „ON“	FP „ON“	Präzision „ON“	Trefferquote „ON“
1	99,8%	22,7%	99,5%	99,8%	77,3%	0,2%	88,6%	77,3%
2	99,7%	4,6%	99,7%	99,7%	95,4%	0,3%	96,1%	95,4%
3	99,9%	5,4%	99,8%	99,9%	94,6%	0,1%	97,4%	96%
4	99,9%	3,9%	99,8%	99,9%	96,1%	0,1%	98,4%	96,1%
5	99,9%	1%	99,9%	99,9%	99,0%	0,1%	99%	99,1%
6	100%	1,4%	100%	100%	98,6%	0%	98,6%	98,9%

Tabelle 16: Erster Durchlauf der Fallstudie - Vorhersagewerte des initialen Trainings mit Kreuzvalidierung

Tagebuchprotokoll zur ersten Iteration

- Tag1: Man hat das Gefühl, als ob der Prototyp nur zum Ausschalten von Beleuchtungen wäre. Es wurde keine richtige Beleuchtungsaktivierung wahrgenommen.
- Tag 2: Das Deaktivieren von Beleuchtungen funktioniert gut, teilweise zu gut. Wir sind heute in im Dunklen Raum gestanden weil der Prototyp die Beleuchtung beim Betreten deaktiviert hat. Das Aktivieren von Beleuchtungen funktioniert im Badezimmer gar nicht.
- Tag 3: Lage unverändert. Ausschalten funktioniert nicht so schlecht, Einschalten funktioniert sehr schlecht. Diese Situation ist sehr frustrierend weil das Einschalten das wichtigere Ereignis ist.
- Tag 4: Keine Neuigkeiten, die Frustration nimmt zu. Wir stehen viel zu oft in einem dunklen Raum weil der Prototyp die Beleuchtung nicht aktiviert oder aktive Beleuchtungen beim Betreten deaktiviert obwohl wir gerne eine aktive Beleuchtung hätten.
- Tag 5: Keine Verbesserung in Sicht.
- Tag 6: Heute hat der Prototyp überraschend gut gearbeitet. Wir sind selten in Dunklen Räumen gestanden obwohl wir gerne Licht gehabt hätten.
- Tag 7: Heute hat die Vorhersage der Beleuchtungsaktivierungen wieder äußerst schlecht funktioniert. Uns ist kein einziges Mal bewusst, bei dem der Prototyp eine Beleuchtung korrekt aktiviert hätte. Die Vorhersagequalität ist für den praktischen Einsatz viel zu schlecht.

Daten zum zweiten Durchlauf der Fallstudie

Modell	TP „OFF“	FP „OFF“	Präzision „OFF“	Trefferquote „OFF“	TP „ON“	FP „ON“	Präzision „ON“	Trefferquote „ON“
1	99,6%	24,9%	98,7%	99,6%	75,1%	0,4%	85,4%	75,1%
2	98,7%	4,9%	98,0%	98,7%	95,1%	1,3%	96,0%	95,1%
3	99,3%	5,4%	98,8%	99,3%	94,6%	0,7%	99,1%	95,8%
4	99,7%	5,1%	99,6%	99,6%	99,7%	0,3%	96,7%	94,9%
5	99,5%	4,3%	99,4%	99,5%	95,7%	0,5%	98,4%	99,3%
6	99,9%	3,8%	99,8%	99,9%	96,2%	0,1%	96,2%	96,9%

Tabelle 17 Zweiter Durchlauf der Fallstudie - Vorhersagewerte des initialen Trainings mit Kreuzvalidierung:

Tagebuchprotokoll zur zweiten Iteration

- Tag1: Im Vergleich zur Vorhergehenden Iteration schein das Aktivieren von Beleuchtungen nun zumindest teilweise zu funktionieren. Heute gab es hier einige Erfolgserlebnisse in dem die Beleuchtung korrekt aktiviert wurde. Es gab aber auch falsche Deaktivierungen der Beleuchtung wie in Iteration 1.
- Tag 2: Die Vorhersageleistung für das Ausschalten von Beleuchtungen funktioniert so wie in der ersten Iteration. Beim Einschalten von Beleuchtungen ist eine Verbesserung zu erkennen. Heute gab es ein paar Situationen in denen beides sehr gut funktioniert hat.
- Tag 3: Die Vorhersageleistung für das Einschalten ist zwar besser geworden, es gibt hier aber dennoch eine relativ hohe Wahrscheinlichkeit, dass es nicht funktioniert. Heute ist es vermehrt wieder dazu gekommen, dass beim Betreten eines Raumes die Beleuchtung deaktiviert wurde obwohl wir die Beleuchtung gerne aktiv hätten.
- Tag 4: Keine Neuigkeiten. Teilweise funktionieren die Vorhersagen einwandfrei. Bei Aktivieren von Beleuchtungen gibt es aber immer wieder Probleme.
- Tag 5: Das Vorhersageverhalten zur Beleuchtungsaktivierung ist zwar besser als in Iteration 1, aber hat teilweise nach wie vor viele Fehler. Heute sind wir wieder oft im Dunkeln gestanden.
- Tag 6: Keine Änderung oder Besonderheiten am Verhalten entdeckt.
- Tag 7: Einige komplett korrekte Vorhersagen für Beleuchtung an und Aus. Aber nach wie vor einige Probleme beim Aktivieren von Beleuchtungen.

Auswertung Beleuchtungsaktivierungen

Beleuchtung	Aktivierung pro Tag (Mittelwert)	
	15.07.2017-31.07.2017	1.11.2017-15.11.2017
Badezimmer Hauptlicht	4,8	9,1
Küche Hauptlicht	2,2	3,2
Wohnbereich – Esstisch	1,8	2,7
Wohnbereich – Fernseher	1,9	3,6
Wohnbereich – Mitte	0,3	0,8
Wohnbereich - Spieleecke	0,5	0,9

Tabelle 18: Auswertung der Beleuchtungsnutzung zu Beginn der Aufzeichnung und zum Ende der Evaluierung

ABKÜRZUNGSVERZEICHNIS

CRISP-DM … Cross Industry Standard Process for Data Mining

TP … True Positive

FP … False Positive

ABBILDUNGSVERZEICHNIS

TABELLENVERZEICHNIS

LITERATURVERZEICHNIS

Augusto , J. C., & Nugent, C. D. (2006). Smart Homes Can Be Smarter. In J. C. Augusto, & C. D. Nugent, *Designing Smart Homes* (S. 1-16). Berlin, Heidelberg: Springer.

Augusto, J., & Nugent, C. (2006). Smart Homes Can Be Smarter. In J. Ausgusto, & C. Nugent, *Designing Smart Homes. Lecture Notes in Computer Science, vol 4008.* Berlin, Heidelberg: Springer.

Azevedo, A., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. *IADIS European Conference Data Mining* (S. 182–185). Amsterdam: IADIS.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*(53), S. 370-418.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Springer.

Bramer , M. (2013). *Principles of Data Mining.* London: Springer.

Breiman, L. (1994). *Bagging Predictors.* University of California, Department of Statistics, Berkeley California.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISPDM 1.0 step-by-step data mining guide.* CRISP-DM.

Cook, D. J. (2012). How Smart Is Your Home? *Science (New York, N.Y), 335*(6076), S. 1579-1581.

Cook, D. J., Crandall, A. S., Thomas, B. L., & Krishnan, N. C. (2013). CASAS: A smart home in a box. *Computer, 46*(7), S. 62-69.

Cook, D. J., Youngblood, M., Heierman, E. O., Gopalratnam, K., Rao, S., Litvin, A., & Khawaja, F. (2003). MavHome: An agent-based smart home. In Pervasive Computing and Communications. *roceedings of the First IEEE International Conference on Pervasive Computing and Communications* (S. 521-524). IEEE.

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*(20).

Das, B., Chen, C., Seelye, A., & Cook, D. (2011). An automated prompting system for smart environments. In B. Abdulrazak, S. Giroux, B. Bouchard, H. Pigot, & M. Mokhtari, *Toward Useful Services for Elderly and People with Disabilities* (S. 9-16). Springer.

Das, S. K., Cook, D. J., Battacharya, A., Herman, E. O., & Lin, T. Y. (2002). The role of prediction algorithms in the MavHome smart home architecture. *IEEE Wireless Communications, 9*(6), S. 77-84.

Demiris, G., Skubic, M., Rantz, M., Keller, J., Aud, M., Hensel B., & He, Z. (2006). Smart home sensors for the elderly: a model for participatory formative evaluation. *human-computer interaction, 6*(7).

Demsar, J. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*(7), S. 1-30.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation, 5*(7), S. 1895-1923.

Dietterich, T. G. (2000). An Experimental Comparison of Three Methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning, 40*(2), S. 139-157.

Dixit, A., & Naik, A. (2014). Use of Prediction Algorithms in Smart Homes. *International Journal of Machine Learning and Computing, 4*(2), S. 157.

Domingos, P., & Pazzini, M. (1997). On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. *Machine Learning*(29), S. 103-130.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics., 7*(1), S. 1-26.

Eisenhardt, K. M. (1989). Building Theories from Case Study Research. *Academy of management review, 14*(4), S. 532-550.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 37-54.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting knowledge from volumes of data. *Communications of the ACM, 39*(11), 27-34.

Flach, P. (2012). *Machine learning: the art and science of algorithms that make sense of data.* Cambridge University Press.

Flick, U. (2010). *Qualitative Sozialforschung. Eine Einführung.* Rowohlt: Reinbek .

Frawley, W., Piatetsky-Shapiro, G., & Matheus, C. (1992). knowledge discovery in databases - an overview. *Ai Magazin, 13*(3), 57-70.

Freund, Y., & Schapire, R. (1999). A Short Introduction to Boosting. *Journal of Japanese Society for Artificial Intelligence, 14*(5), S. 771-780.

Friedman, J., Bentley, J., & Finkel, R. (1977). An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software, 3*(3), S. 209-266.

Gaber, M. M., Zaslavsky, A., & Krishnaswamy, S. (2010). Data Stream Mining. In O. Maimon, & L. Rokach, *Data Mining and Knowledge Discovery Handbook* (S. 759-787). Sprinter.

Gill, K., Yang, S. H., Yao, F., & Lu, X. (2009). A zigbee-based home automation system. *IEEE Transactions on consumer Electronics, 55*(2).

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques.* US: Morgan Kaufmann Publishers Inc.

Hulten, G., Spencer, L., & Domingos, P. (2001). Mining time-changing data streams. *In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining* (S. 97-106). ACM.

Hunt, D. R. (1979). The use of artificial lighting in relation to daylight levels and occupancy. *Building and environment, 14*(1), S. 21-33.

Jakobi, T., Ogonowski, C., Castello, N., Stevens, G., & Wulf, V. (7 2016). Smart Home Experience Journy: Über den Einsatz und die Wahrnehmung von Smart Home-Technologien im Alltag. *Mittelstand-Digital Wissenschaft trifft Praxis*.

Johns, M. (1961). An empirical Bayes approach to nonparametric two-way classification. In H. Solomon, *Studies in item analysis and prediction* (S. 221-232). Palo Alto, CA: Stanford University Press.

Kelly, M. G., Hand, D. J., & Adams, N. M. (1999). The impact of changing populations on classifier performance. *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining* (S. 367-371). ACM.

Kotu, V., & Deshpande, B. (2015). *Predictive analytics and data mining: Concepts and practice with RapidMiner.* US: Morgan Kaufmann Publishers Inc.

Krueger, R. A., & Casey, M. A. (2014). *Focus groups: A practical guide for applied research.* Sage publications.

Love, J. A. (1998). Manual switching patterns in private offices. *International journal of lighting research and technology, 1*, S. 45-50.

Mahdavi, A., Mohammadi, A., Kabir, E., & Lambeva, L. (2008). Shading and Lighting Operation in Office Buildings in Austria: A Study of User Control Behavior. *Building Simulation, 1*(2), S. 111-117.

Marbán, Ó., Mariscal, G., & Segovia, J. (2009). A Data Mining & Knowledge Discovery Process Model. In J. Ponce, & A. Karahoca, *Data Mining and Knowledge Discovery in Real Life Applications.* InTech.

Mitchell, T. M. (1997). *Machine Learning.* New York: McGraw Hill.

Noor, K. (2008). Case Study: A Strategic Research Methodology. *American Journal of Applied Sciences*.

Pan, Y., Liang, J., & Xu, L. (2017). A Study on Intelligent Housekeeper of Smart Home System. *Measuring Technology and Mechatronics Automation (ICMTMA).*

Prati, R. C., Batista, G. E., & Monard, M. C. (2009). Data mining with imbalanced class distributions: concepts and methods. *4th Indian International Conference on Artificial Intelligence (IICAI-09)*, (S. 359-376).

Priti, P., & Yatin, J. (2016). SMART HOME ENVIRONMENT - A BIBLIOMETRIC REVIEW. *International Journal of Recent Scientific Research, 7*.

Pyle, D. (1999). *Data preparation for data mining.* Morgan Kaufmann.

Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning*, S. 81-106.

Quinlan, J. R. (1993). *C4. 5: programs for machine learning.* San Francisco, CA: Morgan Kaufmann.

Reinhart, C., & Wienold, J. (2001). Monitoring user behavior : monitoring and analysis of manual control strategies for lighting and blinds. *International Daylighting : RD & A*, S. 6-7.

Rowley, J. (2002). Using case studies in research. *Management Research News, 25*(1), S. 16-27.

Schiefer, M. (2015). Smart home definition and security threats. *International Conference on IT Security Incident Management & IT Forensics* (S. 114-118). IEEE.

Schulze-Sturm, S. (07 2016). Blickwinkel Smart Home – Studien aus Angebots- und Nachfragesicht. *Mittelstand-Digital Wissenschaft trifft Praxis*.

Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. *Journal Of Data Warehousing, 5*(4), 13-22.

Sixsmith, A., & Johnson, N. (2004). A smart sensor to detect the falls of the elderly. *IEEE Pervasive computing, 3*(2), S. 42-47.

Spring, G., Cook, D., Weeks, D., Dahmen, J., & La Fleur, A. (2017). Analyzing Sensor-Based Time Series Data to Track Changes in Physical Activity during Inpatient Rehabilitation. *Sensors, 17*(10).

Stake, R. E. (1995). *The art of case study research.* Sage.

Strese, H., Seidel, U., Knape, T., & Botthof, A. (2010). *Smart Home in Deutschland.* Berlin: Institut für Innovation und Technik in der VDI/VDE-IT.

Sugiyama, M. (2015). *Introduction to statistical machine learning.* US: Morgan Kaufmann Publishers Inc.

Tapia, E. M., Intille, S. S., & Larson, K. (2004). Activity recognition in the home using simple and ubiquitous sensors. *In Pervasive, 4*, S. 158-175.

timeanddate.de. (2017). Abgerufen am 20. 10 2017 von Feldbach, Steiermark, Österreich — Sonnenaufgang & Sonnenuntergang: https://www.timeanddate.de/sonne/oesterreich/feldbach

Ting, K. M., & Witten, I. H. (1997). *Stacked Generalization: when does it work?*

Tirelli, T., & Pessani, D. (2011). Importance of feature selection in decision-tree and artificial-neural-network ecological applications. Alburnus alburnus alborella: A practical example. *Ecological informatics, 6*(5), S. 309-315.

Weka Java Doc. (17. 10 2017). Von J48: http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/J48.html abgerufen

Witten, I., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann.

Wolpert, D. (1992). Stacked Generalization. *Neural Networks*.

Wolpert, D. H., & Macready, W. G. (1995). *No Free Lunch Theorems for Search.* Santa Fe Institute.

Wu, X., Vipin, K., Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., . . . Steinberg, D. (2007). *Top 10 algorithms in data mining.* London: Springer.

Yin, R. K. (1994). *Case study research: design and methods.* (2 Ausg.). Thousand Oaks: Sage publications.

Yin, R. K. (2013). *Case Study Research: Design and Methods* (5 Ausg.). Sage publications.

Youngblood, G. M., Heierman, E. O., Holder, L. B., & Cook, D. J. (2015). Automation intelligence for the smart environment. *International Joint Conference On Artificial Intelligence* (S. 1513). Lawrence Erlbaum Associates LTD.