Konzeption und Entwicklung einer Datenbasis für Gesundheits-Apps mit qualitätsgeprüften Lebensmitteldaten aus externen Quellen

Von der Fakultät für Medizin und Gesundheitswissenschaften der Carl von Ossietzky Universität Oldenburg zur Erlangung des Grades und Titels eines

Doktor der Ingenieurswissenschaften
(Dr. -Ing.)

angenommene Dissertation

von Herrn Alexander Münzberg

geboren am 24.08.1987 in Bad Bergzabern

Gutachter

Universitätsprofessor Dr. -Ing. Andreas Hein

Weitere Gutachter

Professor Dr. rer. medic. Norbert Rösch

Tag der Disputation: 11.03.2022

Erklärung

Ich erkläre hiermit, dass ich die vorliegende Dissertation selbstständig verfasst und nur die angegebenen Hilfsmittel benutzt habe.

Des Weiteren erkläre ich, dass die Dissertation weder in ihrer Gesamtheit noch in Teilen einer anderen Hochschule zur Begutachtung in einem Promotionsverfahren vorliegt oder vorgelegen hat.

Die Regeln der Carl von Ossietzky Universität Oldenburg zur guten wissenschaftlichen Praxis wurden befolgt.

Im Zusammenhang mit dem Promotionsvorhaben wurden keine kommerziellen Vermittlungs- oder Beratungsdienste (Promotionsberatung) in Anspruch genommen.

Oldenburg, den 01.12.2021

Anmerkung

In der vorliegenden Arbeit wird auf aufgrund der besseren Lesbarkeit das generische Maskulinum verwendet. Weibliche und anderweitige Geschlechteridentitäten sind dabei ausdrücklich mitgemeint.

Danksagung

Im Besonderen danke ich Herrn Universitätsprofessor Dr. -Ing. Andreas Hein und Herrn Professor Dr. rer. medic. Norbert Rösch, für die Unterstützung, Betreuung, sowie deren Ratschläge und Diskussionen während der Durchführung der vorliegenden Arbeit.

Des Weiteren bedanke ich mich

- bei allen Kolleginnen und Kollegen der AG Telehealth, insbesondere der Leitung Herrn Professor Dr. rer. medic. Norbert Rösch sowie Frau Janina Sauer, für deren Mitarbeit an Projekten, an Publikationen und an hilfreichen Diskussionen im Bezug zur Dissertation.
- Bei allen Partnern des DiDiER-Projektkonsortiums, für die gute Zusammenarbeit an Konzepten, für wichtige und hilfreiche Diskussionen und Ratschläge sowie der Mitarbeit an Veröffentlichungen.
- Bei Herrn Dr. -Ing. Christian Lüpkes, Gruppenleiter „Datenmanagement und -analyse für die Versorgungsforschung“ im FuE-Bereich Gesundheit des OFFIS – Institut für Informatik, für die hilfreichen fachlichen Ratschläge im Bereich der Datenanalyse und Datenbanken.
- Bei Frau Lisa Happe, Mitarbeiterin in der Abteilung Assistenzsysteme und Medizintechnik unter der Leitung von Herrn Universitätsprofessor Dr. -Ing. Andreas Hein, für die hilfreiche administrative Unterstützung und Ratschläge bei den Teilnahmen an Seminaren.
- Bei Frau Katharina Münzberg und Frau Janina Sauer, für das Korrekturlesen der Dissertation.

Ein weiterer herzlicher Dank gilt meiner Frau Sonja Münzberg und meinen Eltern, Günter und Gisela Münzberg, für die beständige moralische Unterstützung und Motivation während meiner akademischen Laufbahn.

Meiner Familie

Kurzfassung

Die Zahl der Gesundheitsapps zum Thema Ernährung wächst in den App-Stores der großen Smartphone-Hersteller stetig. Für medizinische und nicht-medizinische Anwendungszwecke in den genannten Apps sind qualitativ hochwertige und gut strukturierte Daten zu Lebensmitteln sowie Lebensmittelprodukten von großer Bedeutung. Da europaweit keine standardisierten Datensätze zu allen Lebensmitteln verfügbar sind, nutzen die Anbieter dieser Gesundheitsapps entweder verschiedene externe Datenquellen oder betreiben eigene Datenbanken, die mit Informationen von Produktherstellern oder Dritten befüllt werden. Das Problem beim Umgang mit solchen Daten ist, dass sie oft von schlechter Datenqualität sind, da Teile der Datensätze entweder unvollständig und fehlerhaft sind oder dieselben Datensätze aus mehreren verschiedenen Quellen nicht miteinander übereinstimmen. Dies wirkt sich auch auf die Qualität der jeweiligen App aus. Für die Akzeptanz von Gesundheitsapps und deren Erfolge bei der Analyse von Lebensmitteldaten ist eine hochwertige Datenqualität von großer Bedeutung. Die vorliegende Arbeit beschäftigt sich mit der Frage, wie sich Lebensmitteldatensätze aus unterschiedlichen Datenquellen mit Hilfe intelligenter computergestützter Methoden zu einer konsistenten Datenstruktur zusammenführen lassen. Darüber hinaus werden die Daten auf Vollständigkeit und Korrektheit überprüft, um sie in Gesundheitsapps sowie für Analysen in digitalen Anwendungen nutzen zu können. Vor diesem Hintergrund werden Verfahren und Methoden der Datenanalyse recherchiert und entwickelt, um die Qualität von Lebensmitteldatensätzen zu optimieren. Im Zuge der Arbeit werden Lebensmittel- und Lebensmitteldatensätze aus Datenquellen verschiedener Anbieter kombiniert und in einem Data Warehouse gesammelt. Dabei ist es notwendig, im Vorfeld die möglichen Anwendungsgebiete der Daten sowie deren Struktur und Metadaten zu analysieren. Mit Hilfe von Data Profiling Methoden und einer speziell entwickelten Ähnlichkeits- und Plausibilitätsanalyse werden die Datensätze automatisch auf ihre Qualität geprüft. Dabei kommen Methoden des maschinellen Lernens und Big Data Technologien zum Einsatz. Zur Evaluierung der eingesetzten Methoden werden geeignete wissenschaftliche Verfahren aus dem Bereich der Daten- und Datenmodell-Analyse eingesetzt. Abschließend erfolgt eine Diskussion der verwendeten Verfahren. Diese Diskussion umfasst einen Ausblick auf die verwendeten Verfahren für weitere Anwendungszwecke in einem allgemeinen Kontext. Der Ablauf der Arbeit basiert auf dem Modell "Cross Industry Standard Process for Data Mining", welches den Standardlebenszyklus eines solchen Projektes beschreibt.

Englische Fassung

Abstract

The number of health apps related to nutrition is growing steadily in the app stores of the major smartphone manufacturers. For medical and non-medical application purposes in the apps mentioned, high-quality and well-structured food and food product data are of high importance. As there are no standardized data sets of all food products available in Germany Europe-wide, the providers of the apps described above either use various external data sources or run their own databases, which they fill with information from product manufacturers or third parties. The problem with dealing with such data is that it is often of poor data quality, as parts of the datasets are either incomplete, incorrect, or the same datasets from several different sources are inconsistent to each other. This also impacts the quality of the respective app. The present work deals with the question, how can food data sets of different data sources be combined into a consistent data structure by using intelligent computer-aided methods? Furthermore, the data are checked for completeness and correctness so that they can be used in health apps and for analysis in digital applications. In view of this, data analysis procedures and methods will be researched and developed to optimize the quality of food datasets. In the course of the work, food and food product datasets from data sources of different providers will be unified and collected in a data warehouse. Thereby, it is necessary to analyze in advance the potential application purposes of the data as well as their structure and metadata. Using data profiling methods and a specially developed similarity and plausibility analysis, the data sets are automatically checked for quality. Machine learning methods and Big Data technology are used for this purpose. Suitable scientific methods from the field of data and data model analysis are used to evaluate the methods employed. Finally, the methods used are discussed. The discussion includes an outlook of the methods used for further application purposes in a general context. The workflow of the thesis is based on the "Cross Industry Standard Process for Data Mining" model, which describes the standard lifecycle of this kind of project.

Inhaltsverzeichnis

Erklärung 1

Anmerkung 2

Danksagung 3

Widmung 4

Kurzfassung 5

Abstract 6

1 Einleitung 10

1.1 Motivation und Problemstellung 10

1.2 Ziel der Arbeit 12

1.3 Aufbau und Ablauf der Arbeit 14

2 Grundlagen und Methoden 18

2.1 Grundlagen der Verwendung von Daten mit Informationen zu Lebensmitteln in Gesundheitsapps 18

2.1.1 Analysen von Gesundheitsapps mit Bezug zu Ernährungsinformationen 18

2.1.2 Mögliche Anwendungszwecke einer Lebensmittel-produktdaten-API 19

2.1.3 Risiken bei der Verwendung von Lebensmitteldaten in Gesundheitsapps 21

2.2 Auf Produktverpackungen angegebene Informationen über Lebensmittelprodukte 23

2.2.1 Global Trade Item Number (GTIN) 23

2.2.2 Angabe von Inhaltsstofflisten 24

2.2.3 Angabe von Nährwerten 25

2.2.4 Angabe für häufig auftretende Allergene 26

2.3 Verwendung von Lebensmittel- und Studiendaten aus dem Projekt „Digitale Dienstleistungen in der Ernährungsberatung“ (DiDiER) 26

2.4 Zur Verfügung stehende Datenquellen 27

2.4.1 Bundeslebensmittelschlüssel (BLS) 28

2.4.2 WikiFood.eu 31

2.4.3 Danone 32

2.4.4 OpenFoodFacts.org 33

2.4.5 FoodRepo.org 34

2.4.6 das-ist-drin.de 35

2.5 Attribute und Datenmerkmale der Lebensmitteldatenquellen 36

2.6 Meta- und Wissensdaten 39

2.7 Datenqualität 40

2.8 Methoden der Datenverarbeitung und Datenanalyse 42

2.8.1 Data Profiling 42

2.8.2 Extract, Transform, Load (ETL) 43

2.8.3 Data Mining 44

2.9 Big Data Anwendungen 47

2.9.1 TimescaleDB 48

2.9.2 Das Apache Spark Framework in der Programmiersprache Python (PySpark) 48

2.10 Methoden der Validierung 49

3 Ergebnisse 52

3.1 Angewendete Entwicklungsumgebungen und Werkzeuge 52

3.2 Anwendung von Data Profiling Methoden 53

3.2.1 Ontologie der Datenattribute 54

3.2.2 Metadatenanalyse 56

3.2.3 Datentypanalyse 57

3.2.4 Musteranalyse in Zeichenketten 59

3.3 Zusammenführung der Lebensmittel-Datenquellen in ein zentrales Data-Warehouse 63

3.3.1 Datenbankschema 63

3.3.2 Generierung der Datenbanktabellen-IDs 66

3.3.3 Entwicklung des ETL-Prozesses 68

3.4 Automatisierte Kategorisierung der Lebensmittelproduktdaten im FDWH 71

3.5 Ähnlichkeits- und Plausibilitätsanalyse 72

3.5.1 Ähnlichkeitsanalyse von Zeichenketten 73

3.5.2 Ähnlichkeitsanalyse von numerischen Werten 75

3.5.3 Bestimmung der Plausibilität und Gesamtähnlichkeit 80

3.5.4 Entscheidung für die Anwendung der beschriebenen Methoden zur Ähnlichkeits- und Plausibilitätsanalyse 84

3.6 Konzipierung einer kontextbasierten Lebensmittel-produktdaten-API 85

3.6.1 Die JSON-API 85

3.6.2 Grafische Benutzerschnittstelle 87

3.6.3 Die Suchfunktionen des FDWH 88

3.7 Performante Bearbeitung großer Datensätze mit Hilfe von Big Data Technologie 89

3.7.1 Entscheidung für die Verwendung von TimescaleDB als Datenbanksystem für FDWH-Datensätze 89
3.7.2 Verwendung des Apache Spark Framework in Python für die performante Datenverarbeitung 91

3.8 Auswertung und Evaluierung 92
3.8.1 Auswertung der durch Data Profiling bereinigten Attributwerte 92
3.8.2 Evaluierung und Auswertung der Ähnlichkeits- und Plausibilitätsanalyse 94

4 Diskussion 99

4.1 Interpretation der Ergebnisse und Ausblick des Produktiveinsatzes der FDWH-Daten und entwickelten Methoden 99
4.1.1 Datenqualitätsverbesserung durch Data Profiling sowie Ähnlichkeits- und Plausibilitätsanalyse 99
4.1.2 Der Einsatz von Big Data Technologie 101

4.2 Anteil der einzelnen Prozessschritte im CRISP-Modell am Gesamtprozess 102

4.3 Kommunikationsstandards und Standardisierung von Daten-Bezeichnern und -Ontologien der Lebensmittelprodukdaten 103

4.4 Einsatz der beschriebenen Methoden für weitere mögliche Anwendungsfelder 103

5 Zusammenfassung 106

Literaturverzeichnis I

Vorabveröffentlichungen IX

Internetverweise X

Abkürzungsverzeichnis XVI

Abbildungsverzeichnis XIX

Tabellenverzeichnis XXI

Anhang XXII

Anhang 1: XXII

Anhang 2: XXIV

Anhang 3: XXV

1 Einleitung

1.1 Motivation und Problemstellung

Viele Krankheitsbilder hängen unmittelbar mit der Nahrungsaufnahme des Menschen zusammen [12]. Im Bereich der medizintechnischen Forschung werden in diesem Zusammenhang immer wieder innovative Ideen beschrieben, welche die Behandlung von ernährungsbezogenen Krankheiten mit Hilfe von Informations- und Kommunikations-Technologie angehen (IKT) [59]. Nach Kröger haben Aspekte wie E-Health und Big Data einen entscheidenden Einfluss auf die Digitalisierung und somit auf die Weiterentwicklung des Gesundheitssystems [42]. Nach Eysenbach (2001) ist E-Health ein aufstrebender Bereich im Schnittpunkt von medizinischer Informatik, öffentlicher Gesundheit und Wirtschaft und bezieht sich auf elektronische Gesundheitsdienste und -informationen, die die über das Internet und andere Informations- bzw. Kommunikations-Technologien bereitgestellt oder verbessert werden [28][57]. mHealth beschreibt nach Gigerenzer et al. (2016) den Einsatz von E-Health-Lösungen auf mobilen Geräten [31]. Mobile Smartphone-Anwendungen (Apps) können dem Patienten bei der Prävention von Erkrankungen durch Ernährungsempfehlungen behilflich sein [63]. Nach Evers-Wölk et al. (2018) wird zwischen sogenannten Gesundheitsapps unterschieden, die eigenständig von beliebigen Endnutzern angewendet werden können und Apps mit medizinischer Zweckbestimmung (Medical Apps), die für medizinisches Fachpersonal entwickelt wurden oder gesetzlich als Medizinprodukt eingestuft werden müssen [27]. In den App-Stores der beiden größten Smartphone-Betriebssystem-Anbieter, Google und Apple, gibt es eine große Anzahl von Gesundheitsapps. Nach Brucker (2019) existieren seit April 2019 weit über 100 000 Apps zum Thema Gesundheit, Fitness oder Wohlbefinden [L5]. Laut einer Umfrage des Digitalverbands Bitkom im Jahr 2017 hatte zu dieser Zeit fast jeder zweite Smartphone-Nutzer eine Gesundheitsapp in Gebrauch. 74 % davon setzten derartige Apps ein, um generell ihre Gesundheit zu verbessern, 51 % hatten Spaß an der Selbstvermessung ihrer Körper und Vitaldaten und 48 % wollten generell mehr über den Stand ihrer Gesundheit wissen. Weitere 39 % der Anwender planten, sich mit App-Unterstützung mehr zu bewegen, 26 % nutzten eine der Apps, um sich gesünder zu ernähren und 17 % wollten die Genesung einer Krankheit fördern [L42]. Verbraucher achten heute mehr denn je auf Inhaltsstoffe, Produktaussagen und Transparenz [L4]. Die Statistiken der 42matters AG (2020) zeigen, dass von insgesamt 1 787 048 Apps im iOS App Store 73 980 Apps unter der Kategorie Health & Fitness zu finden sind [L1]. Von 3 319 794 Apps im Google Play Store gehören 120 877 Apps der Kategorie Food & Drink und 108 994 Apps der Kategorie Health & Fitness an (Stand 11. Juni 2020). Das Thema Ernährung spielt in vielen der Apps eine zentrale Rolle. Beispielsweise werden über die elektronischen Ernährungstagebücher dieser Apps die verzehrten Lebensmittel der Anwender über einen gewissen Zeitraum dokumentiert, um diese anschließend einer

gesundheitsspezifischen Analyse (z. B. die in einem definierten Zeitraum aufgenommenen Kalorien) zu unterziehen [38].

Ein großes Problem vieler Apps mit Bezug zur Ernährung ist die Qualität von Lebensmitteldaten. Die Entwicklung solcher Apps wird durch eine mangelnde Verfügbarkeit von plausiblen Daten mit Informationen zu Inhaltsstoffen und Nährwerten verschiedener auf dem Markt erhältlicher Lebensmittelprodukte erschwert. Solche Daten sind insbesondere dann wichtig, wenn beispielsweise ein Arzt oder Therapeut mit Hilfe elektronischer Ernährungstagebücher einen möglichen Zusammenhang zwischen verspeisten Lebensmitteln und aufgetretenen Symptomen erkennen will [64][65][62][C]. Obwohl Informationen über verwendete Zutaten und Nährwerteangaben auf die Lebensmittelverpackungen aufgedruckt werden müssen, besteht derzeit keine Verpflichtung zur Bereitstellung digitaler Informationen. Deshalb gibt es auch keine zentrale standardisierte Datenbank mit den benötigten Informationen über alle auf dem deutschen Markt erhältlichen Lebensmittelprodukten. Die Entwickler der Apps müssen auf Datenquellen von Dritten oder auf eigene Datensammlungen zurückgreifen [61][4][A]. Es gibt professionelle Anbieter, die eine umfangreiche Sammlung von Lebensmitteldaten in Form einer Anwendungsschnittstelle (API) bereitstellen, die einen elektronischen Zugriff mittels eines vom Endgerät verständlichen Datenformats von den Apps aus auf die Lebensmitteldaten zulassen [24][A]. Derartige Angebote sind dennoch teuer in der Anschaffung und unterliegen Lizenzmodellen, die eine Verwendung in Forschungsprojekten erschweren. Eine Einbeziehung solcher kostspieligen Daten in die Forschung kann dazu führen, dass eine spätere kommerzielle Nutzung der Forschungsergebnisse unprofitabel wird. Des Weiteren erlauben deren Anbieter meist keine langfristige Speicherung der Daten durch den Anwender [14]. Da medizinische Daten in Krankenakten jedoch einer Aufbewahrungspflicht von mindestens zehn Jahren unterliegen, erschwert dieser Sachverhalt die Verwendung von Daten solcher Anbieter für den medizinischen Anwendungsbereich [21][L39]. Viele App-Entwickler greifen deshalb auf frei verfügbare Datenquellen zurück, die neben Einträgen von Produktherstellern oftmals durch Daten von Internet-Communitys und (mittels Crowd-Sourcing-Ansatz) von mehreren Personen zusammengeführt wurden. Dementsprechende Datensätze enthalten oft Inkonsistenzen und sind unvollständig oder fehlerhaft, da sie meist nicht adäquat kontrolliert und überprüft wurden. Eine negative Qualität der Daten wirkt sich auf die Qualität der App aus, die diese Daten verwendet. Deshalb wird eine Verwendung der Daten ohne eine umfangreiche Überprüfung und Qualitätsverbesserung sowie dessen Vervollständigung nicht empfohlen [A]. In vielen Gesundheitsapps wurde auf die Einbindung von Lebensmittelproduktdatenbanken verzichtet. In diesen Apps müssen meistens die Anwender verzehrte Lebensmittel und deren Informationen selbst einpflegen. Dadurch entsteht mitunter eine uneinheitliche Datenstruktur und zwischen den Datenelementen können Inkonsistenzen auftreten. Uneinheitlich strukturierte Daten erschweren es, diese nachträglich zu verarbeiten [46]. In Ballard (2019) wird die Möglichkeit beschrieben, einen Text auf Produktverpackungen via Smartphone zu scannen und

automatisiert zu erkennen und in strukturierte Daten umzuwandeln [L4]. Die vorliegende Arbeit versucht das Problem des Vorhandenseins strukturierter Lebensmitteldaten durch die Zusammenführung mehrerer externer Datenquellen zu lösen, deren Daten intensiv auf Qualität überprüft werden.

1.2 Ziel der Arbeit

Es stellt sich die Frage, ob sich eine Datensammlung von geprüften Lebensmitteldaten erzielen lässt, welche diese Daten von jeweils verfügbaren Datenquellen vereinheitlicht zusammenfasst, um eine große Datenbasis zu erzielen. Über eine API könnten die zusammengefassten Daten an Apps mit einem medizinischen bzw. gesundheitsspezifischen Kontext geliefert werden. Damit die Daten für die jeweiligen Apps von Nutzen sind, müssen sowohl Inkonsistenzen als auch Fehler in den Daten erkannt und bereinigt sowie Datensätze vervollständigt werden.

Das Ziel dieser Arbeit ist, die verfügbaren Datenquellen zu nutzen und die Qualität der Daten dahingehend zu verbessern, dass diese in der Forschung für die Entwicklung von Gesundheitsapps herangezogen werden können. Hierbei wird eine Methode entwickelt, die das Sammeln, Analysieren und Korrigieren von Daten aus vorhandenen Quellen ermöglicht. Ein weiterer Schritt ist die Entwicklung der zugehörigen API. Durch diese sollen weitestgehend nur konsistente und korrekte Daten weitergegeben werden. Datensätze, die unvollständig sind und sich nicht durch die im weiteren Verlauf der Arbeit beschriebenen Methoden vervollständigen lassen, werden nicht direkt aus dem Datenbestand der API entfernt. Ist ein Teil der Informationen des unvollständigen Datensatzes für einen bestimmten Anwendungskontext nützlich, kann dieser mit einem Hinweis auf die fehlenden Datenbereiche von der API gesendet werden. Somit ist der Informationsumfang der API sowie deren Nützlichkeit höher. Zum Beispiel benötigt eine App mit einem elektronischen Ernährungstagebuch Informationen über die Inhaltsstoffe eines bestimmten Müsliriegels. Im Falle dieses Szenarios wird bezüglich einer Lebensmittelallergie des Anwenders die Information benötigt, ob in dem Müsliriegel Haselnüsse enthalten sind. Da der Kontext in dem Szenario auf dem Inhaltsstoff Haselnuss liegt, ist es beispielsweise irrelevant, ob in dem Müsliriegel Laktose enthalten ist. Ist nun die Information über den Gehalt von Laktose in dem Datensatz des Müsliriegels nicht vorhanden, ist der Datensatz somit unvollständig, ist aber die Information über den Inhaltsstoff Haselnuss vorhanden, so kann die API den Datensatz mit der benötigten Information weitergeben. Zusätzlich wird ein Hinweis über die Unvollständigkeit des Datensatzes, in diesem Fall über den fehlenden Laktose-Gehalt, weitergegeben.

Zur Zusammenführung der Daten aus den verschiedenen Datenquellen mit Informationen über Lebensmittelzusammensetzungen und Lebensmittelproduktdatenbanken sowie der Qualitätsverbesserung der Daten, werden weitere Zwischenziele definiert. Die Datensätze der

verschiedenen Lebensmitteldatenquellen werden mit Hilfe des Extract-, Transform-, Load-Prozesses (ETL-Prozess) aus den Datenquellen extrahiert, in eine einheitliche Datenstruktur und ein einheitliches Datenformat transformiert und in ein zentrales Data Warehouse (im Folgenden FDWH für „Food Data Warehouse“ genannt) geladen. Somit liegen alle Daten in einem einheitlichen Schema vor und können weiteren Verfahren der Qualitätsverbesserung unterzogen werden. Inkonsistenzen in den Datensätzen werden mit Hilfe von Data Profiling Methoden aufgedeckt und bereinigt [A][C]. Durch die Konzeption und Entwicklung einer Ähnlichkeits- und Plausibilitätsanalyse werden Fehler in Datensätzen erkannt und unvollständige Datensätze weitestgehend vervollständigt. Sind sich zwei Produkte ähnlich und bei einem der Produkte fehlen Informationen im Datensatz, so können diese gegebenenfalls von dem jeweils anderen der beiden Produkte bezogen werden. Zum Beispiel haben Produkt *A*, eine koffeinhaltige Limonade von Hersteller *X* sowie Produkt *B*, eine koffeinhaltige Limonade von Hersteller *Y* jeweils die gleichen Inhaltsstoffe, aber bei Produkt *B* fehlen die Nährwerteinformationen. Durch die Ähnlichkeitsanalyse wird festgestellt, dass sich die beiden Produkte im Bereich von Produktnamen und Produktkategorie sowie im Bereich der Inhaltsstoffe ähnlich sind. Somit kann Produkt *B* gegebenenfalls die Nährwerteinformationen von Produkt *A* übernehmen. Mittels der eigens entwickelten Plausibilitätsanalyse werden die Datensätze nach der Vervollständigung auf ihre Plausibilität hin untersucht, damit anschließend eine Behebung erkannter Fehler oder Inkonsistenzen stattfinden kann. Für die konzipierten Methoden zur automatisierten Datenqualitätsoptimierung (sowohl Data Profiling Methoden als auch die Ähnlichkeits- und Plausibilitätsanalyse) werden verschiedene, mitunter auch eigens speziell dazu entwickelte Verfahren, ausgewählt (darunter fallen Data Mining und Text Mining Methoden bzw. Methoden des maschinellen Lernens sowie bereits existierende Methoden des Data Profilings) und angewandt.

Im weiteren Verlauf der Arbeit findet eine Auswertung der durch Data Profiling bereinigten Datensätze statt. Zudem werden die eingesetzte Ähnlichkeits- und Plausibilitätsanalyse mithilfe der Bildung von Wahrheitsmatrizen und der Auswertung durch statistische Methoden, die in Datenanalyseprojekten üblich sind, evaluiert. Bei stetig wachsender Anzahl von sich über die Zeit veränderlichen Daten und deren Analyse kommt es vor, dass eine Verarbeitung mit herkömmlichen Datenverarbeitungsmethoden und -Werkzeugen in Bezug zur Rechenperformance an ihre Grenzen kommt. Dieser Sachverhalt wurde auch innerhalb der vorliegenden Arbeit beobachtet. In diesem Kontext wird eine Betrachtung von Big Data Lösungen durchgeführt. Aus einer Auswahl von geeigneten Big Data Lösungen, wird die bestgeeignete Lösung für die Speicherung der Lebensmittelproduktdaten sowie die durchzuführenden Datenverarbeitungsprozesse zur Qualitätsoptimierung eingesetzt.

Zusammengefasst sind Ziele der vorliegenden Arbeit die Konzeption und Entwicklung der Lebensmitteldaten- und Lebensmittelproduktdaten-API sowie der zentralen Datenbasis (das FDWH) und aller zur Performanz- und zur Datenqualitätsoptimierung beitragenden Verfahren und Methoden inklusive deren Evaluierung.

In der nachfolgenden Auflistung sind die Ziele der Arbeit (aufgeteilt in Primärziel und dafür durchzuführende Zwischenziele) punktuell aufgeführt:

- Primärziel: Entwicklung einer API, die qualitativ hochwertige Lebensmittel- und Lebensmittelproduktdaten von frei verfügbaren Datenquellen an medizinische Apps und Gesundheitsapps liefert, unter Berücksichtigung des jeweiligen Anwendungskontextes.
- Zwischenziele:
  - Zusammenführung von Datenquellen mit Hilfe des ETL-Prozesses in ein zentrales Data Warehouse.
  - Inkonsistenz-Erkennung und -Bereinigung der Daten im Data Warehouse durch Data Profiling Methoden.
  - Vervollständigung der Daten durch Entwicklung einer Ähnlichkeitsanalyse und Fehlererkennung bzw. -bereinigung durch Entwicklung einer Plausibilitätsanalyse
    - unter Anwendung von Data Mining und Text Mining Methoden sowie Methoden des maschinellen Lernens.
  - Automatisierung der Prozesse zur Datenqualitätsverbesserung der Lebensmittel- und Lebensmittelproduktdaten
  - Auswahl und Anwendung von passenden Big Data Technologien zur performanten Datenspeicherung und -verarbeitung.

1.3 Aufbau und Ablauf der Arbeit

Wie es für viele Projekte mit dem Bezug zur Datenverarbeitung und Datenanalyse üblich ist, orientieren sich Aufbau und Ablauf der Arbeit am Cross Industry Standard for Data Mining Vorgehensmodell (kurz: CRISP-DM) [D]. Dieses standardisierte Modell wurde durch ein Konsortium aus den Firmen NCR Corporation, Daimler AG, SPSS, Teradata und OHRA entwickelt [15][18]. Es beschreibt abstrakt die verschiedenen durchzuführenden Schritte während des Lebenszyklus eines Data Mining Projektes. Abbildung 1.1 zeigt eine grafische Darstellung des CRISP-DM-Modells. Wie an den Pfeilen innerhalb der Darstellung zu erkennen ist, werden einige der Prozessschritte im Lebenszyklus wiederholt durchgeführt, sofern dies einer der Prozessteile im Modell erfordert. Nachfolgend werden die einzelnen Prozessschritte dieses Modells genauer erläutert [18].

- Aufgabenverständnis (Business Understanding):
  - Ein grundlegendes Verständnis des jeweiligen Fachgebietes wird gewonnen. Ist-Zustand und zur Verfügung stehende Ressourcen werden ermittelt. Des Weiteren werden die Kriterien für den Erfolg des Projektes festgelegt.
- Datenverständnis (Data Understanding):

- Es stellt sich die Frage, welche Informationen die zur Verfügung stehenden Daten beinhalten und welche für die Erreichung des Projektziels benötigt werden. In diesem Prozessschritt wird die genaue Bedeutung der Daten analysiert.

- Datenvorbereitung (Data Preparation):
  - Die Daten werden hier auf den im nächsten Punkt erklärten Modelling-Prozessschritt vorbereitet. Daten werden selektiert, in geeignete und einheitliche Strukturen gebracht, zusammengeführt, auf Konsistenz sowie auf Vollständigkeit geprüft und gegebenenfalls korrigiert.
- Modellbildung (Modelling):
  - Hiermit beginnt die eigentliche Datenanalyse und somit das eigentliche Data Mining (siehe Abschnitt 2.8.3). Es werden geeignete Analyse- und Auswertungsverfahren ausgewählt, um nützliche Informationen aus den Daten zu erhalten.
- Evaluation:
  - In diesem Abschnitt werden alle verwendeten Analyseverfahren und deren Ergebnisse auf Richtigkeit und Gültigkeit geprüft und bewertet.
- Produktiveinsatz (Deployment):
  - Die Daten und Modelle werden auf ihren Produktiveinsatz hin überprüft und vorbereitet.

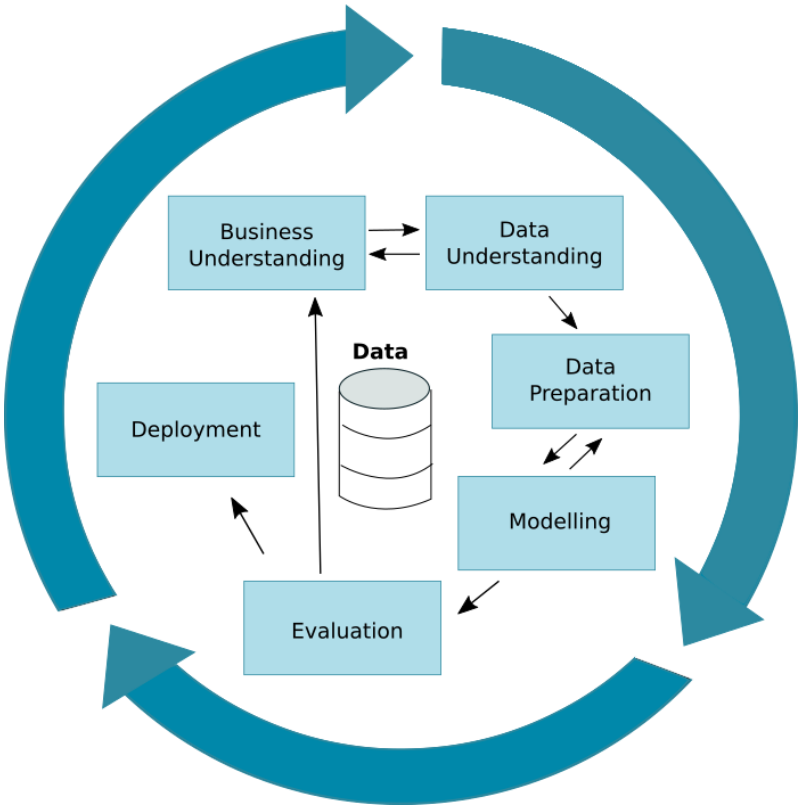


Abbildung 1.1: CRISP-DM-Modell (in Anlehnung an Chapman et al. (2000) [15] und Cleve & Lämmel (2020) [18])

Die zu Prozessschritt 1 (Business Understanding) des CRISP-DM-Modells zugehörigen Teilschritte der vorliegenden Arbeit wurden in gewissem Umfang bereits im einleitenden Kapitel erläutert. Es wurde das Problem beschrieben und Ziele für den Projekterfolg festgelegt. Ein weiterer Teil der Arbeit, der sich diesem Prozessschritt zuordnet, beschäftigt sich mit der Qualität und Quantität von Lebensmittel- und Lebensmittelproduktdaten in bereits existierenden Gesundheitsapps.

Kapitel 2 befasst sich insgesamt mit den Grundlagen und Methoden der Arbeit, dessen Abschnitte wie folgt aufgeteilt sind und sich mit Teilschritten des CRISP-DM-Modells in Verbindung bringen lassen. Das Kapitel behandelt potenzielle Anwendungszwecke sowie das zwischen 2016 und 2019 durchgeführte, vom Bundesministerium für Bildung und Forschung (BMBF) geförderte Projekt „Digitale Dienstleistungen in der Ernährungsberatung“ (DiDiER) mit Bezug zu Lebensmittel- und Lebensmittelproduktdaten als wichtige Grundlagen für die weiteren Projektschritte. Des Weiteren werden die zur Verfügung stehenden Ressourcen erläutert. Diesbezüglich werden neben den externen Datenquellen als Lieferanten für Gesundheitsdaten und deren Beschreibungen auch weitere Wissens- und Metadaten beschrieben. Hierbei wird auf das Verständnis der zur Verfügung stehenden Daten eingegangen, welches durch Schritt 2 (Data Understanding) im CRISP-DM-Modell repräsentiert wird. Im weiteren Verlauf wird der Begriff Datenqualität und die damit verbundenen Merkmale, auf die sich die nachfolgend beschriebenen Prozessschritte Data Preparation und Modelling beziehen, erläutert. Anschließend werden die Analyse- und Datenverarbeitungs-Methoden dargestellt, die in der Arbeit eingesetzt werden und sich auf Prozessschritt 3 (Data Preparation) und Prozessschritt 4 (Modelling) des CRISP-DM-Modells beziehen. Zuletzt werden die Methoden beschrieben, die für die Validierung von Analyseverfahren zum Einsatz kamen und die sich auf den CRISP-DM-Prozessschritt 5 (Evaluation) beziehen.

Kapitel 3 beschreibt die Ergebnisse der vorliegenden Arbeit. Zunächst werden die in der vorliegenden Arbeit zum Einsatz gebrachten Entwicklungsumgebungen und Werkzeuge beschrieben. Die Datenvorverarbeitung (CRISP-DM-Prozessschritt 3, Data Preparation) geschieht mithilfe sogenannter Data Profiling Methoden und deren Regeln sowie der Transformation während des ETL-Prozesses. Des Weiteren werden während der Datenvorbereitung Lebensmittelkategorien mit den passenden Lebensmitteldatensätzen verknüpft. Während dieses Prozessschritts wird das Food Data Warehouse (FDWH) aufgebaut und verschiedene Methoden zur Verbesserung und Beurteilung der Datenqualität angewendet. Zusätzlich werden, in Anlehnung zu Prozessschritt 4 (Modelling) des CRISP-DM-Modells, eigens entwickelte Modelle und Methoden zur Ähnlichkeits- und Plausibilitätsanalyse der Lebensmittelproduktdaten sowie die API zur Lieferung der Lebensmittelproduktdaten im FDWH an Gesundheitsapps beschrieben. Die nachfolgend vorgestellten und in der Arbeit angewandten Big Data Technologien werden zur performanten Verarbeitung großer Datensätze eingesetzt. In Bezug zu Schritt 5 des Modells (Evaluation) finden nachfolgend Überprüfungen und Bewertungen der Gültigkeit der

angewandten Methoden zur Qualitätsverbesserung der Lebensmittel- und Lebensmittelproduktdatensätze sowie deren Ergebnisse statt.

Ein Ausblick auf Produktiveinsatz der Daten des FDWH und der entwickelten Methoden zur Qualitätsverbesserung (bezüglich CRISP-DM-Prozessschritt 6, Deployment) findet in Kapitel 4 statt. Zusätzlich befasst sich dieses Kapitel mit der Diskussion über die Interpretation der Ergebnisse der Arbeit. Weitere Diskussionen befassen sich mit dem Anteil der Prozessschritte im CRISP-DM-Modell am Gesamtprozess, der Standardisierung von Zutatenbezeichnern und Ontologien der Lebensmittelproduktdaten für eine vereinfachte Verarbeitung in digitalen Anwendungen und dem Einsatz der beschriebenen Methoden im allgemeinen Kontext. Abschließend liefert Kapitel 5 eine Zusammenfassung der gesamten Arbeit.

2 Grundlagen und Methoden

2.1 Grundlagen der Verwendung von Daten mit Informationen zu Lebensmitteln in Gesundheitsapps

2.1.1 Analysen von Gesundheitsapps mit Bezug zu Ernährungsinformationen

Um einen Einblick zu erhalten, wie derzeitige Gesundheitsapps in den App-Stores der beiden großen Smartphone-Hersteller Apple (der Firma Apple Inc.) und Google (der Firma Google LLC) Lebensmittel- und Lebensmittelproduktdatenbanken einbinden, werden im Zuge des in Abschnitt 2.3 beschriebenen Projektes DiDiER 16 verschiedene Apps mit Ernährungstagebüchern und Produktinformationen analysiert. In einigen der Apps sind Daten aus Datenbanken mit Informationen über Lebensmittelzusammensetzungen und Lebensmittelproduktdatenbanken eingebunden. Manche Apps binden keine Informationen aus diesen Datenbanken ein, sondern lassen diese durch den Anwender der App eingeben. Die Analyse bringt die Feststellung mit sich, dass sowohl Anzahl auch als die Informationsdichte (Menge an Informationen, die zu den genannten Attributen aus Tabelle 2.1 enthalten sind) in den meisten Apps mangelhaft ist. Nur drei davon bieten zufriedenstellende Treffer und Informationen bei der Produktsuche. Dies sind die Apps Ernährungstagebuch von My Daily Bits LLC zur Überwachung von Essverhalten [L38], dem Kalorienzähler von YAZIO GmbH [L63] und die Lebensmittel und Cosmetic Scanner-App der Firma Codecheck AG als Einkaufsassistent für Lebensmittelproduktdaten mit Barcodescanner [L11]. Die Informationen der beiden erstgenannten Apps können nur über ein kostenpflichtiges Abo bezogen werden. Nur die Entwickler der App der Firma Codecheck AG bieten zusätzlich einen Zugriff auf die in der getesteten App verwendeten Lebensmittelproduktdatenbank an, der aber kostenpflichtig ist. Eine Tabelle mit Details zum durchgeführten Test der Gesundheitsapps befindet sich im Anhang (Anhang 1).

In Pirrung (2020) werden weitere Gesundheitsapps mit Bezug zu Nahrungsmittelunverträglichkeiten analysiert [56]. Einbezogen sind die Apps Nutrimizer, MySymptoms, CaraCare, MyFoodDB und Foody. Hierbei überzeugt nur die App MyFoodDB mit einem positiven Kosten- und Nutzverhältnis sowie einer ausreichend guten Datenanbindung zu Lebensmitteldaten. Den vollständigen Funktionsumfang erhält man aber nur in der kostenpflichtigen Version der App [56].

Die mangelhafte Integration von umfangreichen Lebensmittelinformationen in den meisten der getesteten Apps zeigt, dass es schwierig ist, an eine qualitativ hochwertige Datensammlung der in Deutschland erhältlichen Produkte zu kommen. Solche Datensammlungen werden über hochpreisige Lizenzmodelle angeboten, wodurch der Preis der Apps steigt, welche die Daten verwenden. Die vom öffentlichen Forschungszentrum (CRP) Henri Tudor in Luxemburg (seit 2015 das Luxemburg Institute

of Science and Technology, abgekürzt: LIST) entwickelte Plattform WikiFood.eu hatte den Ansatz, eine eigens entwickelte Lebensmittelproduktdatenbank durch das Wiki-Prinzip zu füllen. Eine Community von Freiwilligen fügte Daten ein und korrigierte diese gegenseitig [59][64]. Die Plattform steht aber derzeit nicht mehr zur Verfügung, hat aber Daten für die Forschung an der Hochschule Kaiserslautern (in Kooperation mit der Universität Oldenburg) zur Verfügung gestellt. Im Verlauf der vorliegenden Arbeit wird versucht, anhand dieser und anderen frei zur Verfügung stehenden Datenquellen eine Datenbasis zu entwickeln, die Daten für Forschungszwecke in einer akzeptablen Qualität zur Verfügung stellt.

2.1.2 Mögliche Anwendungszwecke einer Lebensmittel-produktdaten-API

Das FDWH dient als Datenlieferant für unterschiedliche Gesundheitsapps. Diesbezüglich werden die Produktdaten an Apps mit medizinischem Bezug, beispielsweise zur Symptombekämpfung, geliefert. Aber auch eine Nutzung der Daten in nicht medizinischen Apps, beispielsweise Apps mit Bezug zu gesunder Ernährung oder Fitness (z. B. durch die Bereitstellung elektronischer Ernährungstagebücher) ist möglich. In diesem Kapitel wird analysiert, welche gesundheitlichen Beeinträchtigungen und Krankheitssymptome in Abhängigkeit mit einzelnen Lebensmittelbestandteilen stehen. Dadurch wird die Erkenntnis erzielt, welche potenziell notwendigen Attribute der jeweiligen Datensätze und Informationen bereitgestellt werden müssen, die aus Lebensmitteldaten unterschiedlicher Datenquellen stammen. In Tabelle 2.1 ist das Ergebnis dieser Analyse dargestellt. Eine Beschreibung der in der Tabelle aufgeführten Symptome und Gesundheitsbeeinträchtigungen in der Ernährungsmedizin erfolgt in Biesalski et al. (2020), in Barth & Kraft (2009) sowie in Rösch (2010) [12][7][59].

Beeinträchtigung/ Symptom	**Beschreibung**	**Abhängige Attribute und Informationen**
Untergewicht	Oft durch Mehrfach-erkrankungen hervorgerufen. Unterversorgung von essenziellen Nährstoffen.	Makronährstoffe, insbesondere Energie, Mikronährstoffe (Vitamine, Mineralstoffe)
Unterernährung (Malnutrition, Frailty/Gebrechlichkeit)	Inadäquate Energie- und Nährstoffzufuhr.	Makronährstoffe, insbesondere Energie und Proteine (bei Protein-Malnutrition),

		Mikronährstoffe (Vitamine, Mineralstoffe).
Essstörungen (Anorexia nervosa, Bulimie nervosa)	Magersucht und Bulimie. Charakteristisch sind extremes Untergewicht und extreme Kontrolle des Essverhaltens.	Makronährstoffe, insbesondere Energie, Mikronährstoffe (Vitamine, Mineralstoffe)
Adipositas (und damit assoziierte Erkrankungen wie Hypertonie, Typ-2-Diabetes, koronare Herzkrankheiten, Herzinsuffizienz, Mamma- und Kolonkarzinom, orthopädische und psychosoziale Probleme)	Adipositas wird umgangssprachlich als Fettsucht bezeichnet. Volumen der Fettzellen nimmt zu und das Fettgewebe im Bauch zeigt besonders hohe Umsatzraten.	Makro- und Mikronährstoffe, insbesondere Energie, Cholesterin, Fett, Protein, Kohlenhydrate, Ballaststoffe, Salz, Calcium und Alkohol.
Diabetes mellitus mit assoziierten Störungen des Kohlenhydrat- und Fettstoffwechsels	Umfasst alle Formen der akuten oder chronischen Hyperglykämie (zu hoher Blutzuckerspiegel, Überzuckerung. Folgeerkrankungen sind vorwiegend durch Gefäßveränderungen geprägt.	Makronährstoffe, insbesondere Energie, Kohlenhydrate, Zucker, Fett, Proteine, Alkohol. Inhaltsstoffliste der Lebensmittelprodukte, um potenziell gefährliche Inhaltsstoffe zu erkennen (z. B. Zucker).
Fettstoffwechselstörungen (Hyperlipidämie)	Genetisch bedingt oder als Folge von Erkrankungen oder Medikamenteneinnahme. Ursache ist oft Überernährung. Behandlung durch lipidsenkende Kost.	Makronährstoffe, insbesondere Fett, Cholesterin, Kohlenhydrate und Ballaststoffe
Nahrungsmittelunverträglichkeiten / Allergien	Allergien bzw. Unverträglichkeiten in	

	Verbindung mit dem Verzehr von Lebensmittel-inhaltsstoffen und auftretenden Symptomen, beispielsweise Lippenschwellungen, Magen-Darm-Beschwerden etc. (in Anlehnung an Rösch (2010) [59]).	Inhaltsstoffliste der Lebensmittelprodukte, um potenziell gefährliche Inhaltsstoffe zu erkennen. Eine weitere wichtige Informationsquelle ist die Pflichtangabe der 14 kennzeichnungspflichtigen Allergene in Lebensmittelprodukten (siehe Abschnitt 2.5.4).

Tabelle 2.1 Gesundheitliche Beeinträchtigungen und Krankheitssymptome in Abhängigkeit mit Lebensmittelbestandteilen

Durch die oben dargestellte Tabelle sowie die zugrunde liegende Literatur wird ersichtlich, dass für die meisten Symptome und Beeinträchtigungen Informationen über Makro- und Mikronährstoffe von Bedeutung sind, insbesondere Kohlenhydrate, Fette und Proteine. Informationen über Inhaltsstoffbezeichner spielen im Falle von Nahrungsmittelallergien und Unverträglichkeiten eine große Rolle.

2.1.3 Risiken bei der Verwendung von Lebensmitteldaten in Gesundheitsapps

Bei der Entwicklung von Gesundheitsapps wird zwischen Apps mit der Zertifizierung „Communauté Européenne“ (CE-Zertifizierung) [L57] und nicht zertifizierten allgemeinen Gesundheitsapps unterschieden. CE-zertifizierte Software unterliegt einer medizinischen Zweckbestimmung und muss bestimmten Anforderungen nach der Medizinprodukteverordnung (EU) 2017/745 (MDR) entsprechen [L14][64].

Bei der Einstufung einer Gesundheitsapp als Medizinprodukt muss unter anderem das Risikomanagement des Produktes nach ISO 14972 sichergestellt werden [L8]. Dies beinhaltet die Identifizierung von Gefährdungen und Gefahren im Zusammenhang mit dem Medizinprodukt für Patienten oder für Anwender und die Einschätzung der Eintrittswahrscheinlichkeit von möglichen Folgen eines Risikos [L58][L26].

Die Verwendung der in dieser Arbeit beschriebenen, externen Lebensmittelproduktdaten kann ein mögliches Gesundheitsrisiko darstellen, je nach Anwendungszweck der Gesundheitsapp, welche die Daten verwendet [L58][L26]. Insbesondere würden Daten mit fehlerhaften Informationen ein erhöhtes Gesundheitsrisiko für Patienten darstellen. In Abhängigkeit des Anwendungszwecks bzw. der Beeinträchtigung eines Patienten, werden fehlerhafte Informationen innerhalb der Attribute eines Lebensmittelproduktdatensatzes die einen Bestandteil der Zusammensetzung eines Lebensmittelproduktes repräsentieren, unterschiedlich gewichtet. Es besteht die Möglichkeit, dass der Verzehr eines Produktes, das aufgrund falscher oder fehlender Informationen von einem Patienten als unbedenklich eingestuft wird, sich für diesen als lebensbedrohlich herausstellt. Des Weiteren gibt es Risiken, die im Falle der Fehl- oder Falschinformierung zwar bei geringem Verzehr keine Auswirkungen auf die Gesundheit eines Patienten haben, sich aber bei vielfachem Verzehr als gesundheitsschädlich herausstellen. Nachfolgend sind Beispiele möglicher Risiken aufgeführt, die im Falle von fehlerhaften Informationen, in einem Bereich des Datensatzes der Informationen über die Zusammensetzung des Produktes liefert, auftreten können.

- Fehlende Zutat in der Inhaltsstoffliste bzw. falsche Information bezüglich der Angabe des Nicht-Vorhandenseins eines der kennzeichnungspflichtigen Allergene (Information, dass das Allergen nicht im Lebensmittelprodukt enthalten ist, obwohl es enthalten ist):
  - Gesundheitsrisiko im Falle eines Patienten mit Allergie bezüglich der betroffenen Zutat oder des betroffenen Allergens.
  - Wegen fehlender Information über das Vorhandensein der Zutat bzw. wegen falscher Information über das Vorhandensein eines kennzeichnungspflichtigen Allergens im Lebensmittelprodukt ist es möglich, dass der Patient das Produkt verzehrt und diesbezüglich an Allergiesymptomen leidet. Schwerstmögliche Reaktion: Anaphylaktischer Schock mit Todesfolge [59][79].
- Falsche Information über die Menge eines der Nährwerte Fett, Kohlenhydrate oder Proteine:
  - Nährwertangaben geringer als in der Realität:
    - Mögliche Gesundheitsrisiken beispielsweise im Falle der Beeinträchtigungen Adipositas, Diabetes mellitius und Fettstoffwechselstörungen wegen zu hoher Nährwertbelastungen aufgrund der Falschinformation [12][47].
  - Nährwertangaben höher als in der Realität:
    - Mögliche Gesundheitsrisiken beispielsweise im Falle der Beeinträchtigungen Untergewicht, Unterernährung und Essstörungen wegen zu geringer Nährwertbelastungen aufgrund der Falschinformationen [12][47].

Die Analyse aller möglichen Risiken im Zusammenhang mit allen Attributbereichen der Lebensmittelproduktdatensätze, für alle möglichen Anwendungszwecke, gestaltet sich als schwierig. Inwiefern die Daten während ihrer Verwendung in der jeweiligen App im Detail genutzt werden, hängt

von der jeweiligen Gesundheitsapp ab. Im Falle einer CE-Zertifizierung einer solchen App, welche die Lebensmittelproduktdaten integriert, müssten diese Risiken detailliert (für den jeweiligen Anwendungszweck) während des Risikomanagementprozesses untersucht werden. Dies gilt insbesondere dann, insofern Qualitätsmängel mit den Folgen der Fehl- bzw. Falschinformation nicht ausgeschlossen werden können [L58][L26].

2.2 Auf Produktverpackungen angegebene Informationen über Lebensmittelprodukte

2.2.1 Global Trade Item Number (GTIN)

Die GTIN (Global Trade Item Number), ehemals EAN-13 (European Article Number mit 13 Ziffern), wird meist in Form eines eindimensionalen Barcodes auf Lebensmittelverpackungen aufgedruckt. Als Beispiel ist in Abbildung 2.1 der eindimensionale Barcode des Produktes Steinofen-Pizza der Marke Wagner aufgeführt. Die Nummer besteht aus 13 numerischen Zeichen, wird von der GS1 Germany GmbH verwaltet und wird zur Identifikation von Handelswaren in über 120 Ländern eingesetzt [34][59]. Die privatrechtlich aufgestellte GS1 Germany GmbH vertritt überwiegend Interessen von Nahrungsmittel-Hersteller und -Händlern [34][59]. Nach Rösch (2010) werden identisch benannte Lebensmittel eines Herstellers mit verschiedenen länderspezifischen GTIN-Barcodes versehen und teilweise sind auch Rezepturen von Produkten an den jeweiligen landestypischen Geschmack angepasst [59]. Manche, meist unverpackte, Lebensmittel sind in Supermärkten mit marktspezifischen Barcodes versehen, die vom Markt frei vergeben werden können. Diese Barcodes können je nach Markt und Verkäufer unterschiedlich ausfallen. Eine Veränderung der Zusammensetzung eines Produktes erwirkt nicht zwingend eine Veränderung des Barcodes [59]. Aus diesen Gründen wird im FDWH mit einem eigenen Schema zur Identifikation der Datensätze von jeweiligen Produkten gearbeitet. Außerdem werden den Datensätzen genaue Zeitstempel zugeordnet, die sich bei Veränderungen des Datensatzes anpassen (siehe Abschnitt 3.3). Das GTIN-Verfahren bietet dennoch die Möglichkeit, beispielsweise mit einer Smartphone-Kamera den zugehörigen Barcode abzuscannen und damit den Produktnamen aus dem FDWH zu selektieren.

Laut der GS1 Germany GmbH wird die GTIN aus der Basisnummer ihrer individuellen globalen Lokationsnummer (GLN), einer von ihnen selbst vergebenen Ziffernfolge sowie einer Prüfziffer gebildet. Die GLN ist ein 13-stelliger Schlüssel, der den globalen Geschäftspartner der GS1 Germany GmbH eindeutig identifiziert [L20].

Abbildung 2.1 Beispiel eines eindimensionalen Barcodes des Produktes Steinofen-Pizza der Marke Wagner

2.2.2 Angabe von Inhaltsstofflisten

Die EU-Verordnung Nr. 1169/2011 [L60] regelt die Angabe von Bezeichnungen für Zutaten, Zusatzstoffe und Aromen Pflicht [L59]. Nach der Verordnung erfolgt die Angabe der Zutaten in absteigender Reihenfolge. Auch die Einzelbestandteile von zusammengesetzten Zutaten einer Inhaltsstoffliste müssen angegeben werden, sofern die Zutat zu mindestens zwei Prozent im Endprodukt enthalten ist. Besteht ein Lebensmittelprodukt nur aus einer einzigen Zutat, so muss keine Inhaltsstoffliste auf die Produktverpackung aufgedruckt werden. Anstelle von Zusatzstoffbezeichnungen kann auch die zugehörige E-Nummer, die die jeweiligen Zusatzstoffe codiert darstellt (z. B. E251 für Natriumnitrat), angegeben werden [L32]. Nachfolgend ist in Abbildung 2.2 die Inhaltsstoffliste des Produktes Steinofen-Pizza der Marke Wagner als Beispiel aufgeführt.

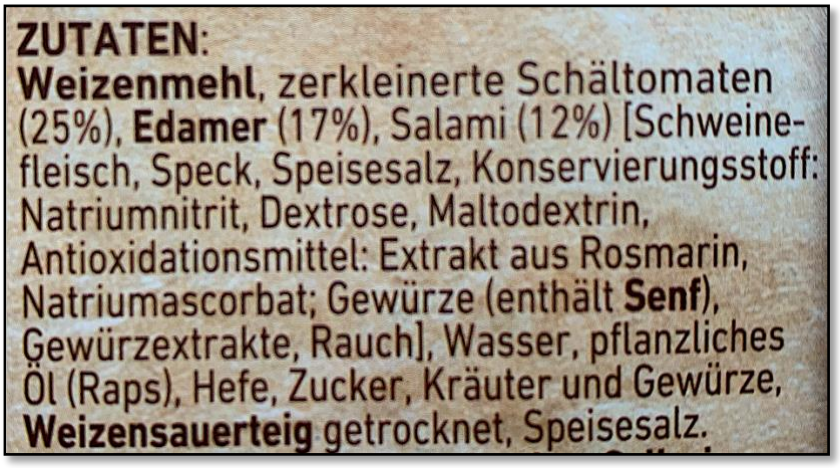


Abbildung 2.2 Beispiel einer Inhaltsstoffliste des Produktes Steinofen-Pizza der Marke Wagner

Die in Abbildung 2.2 gezeigte Inhaltsstoffliste enthält zusätzliche Prozentangaben zur Menge der Zutat im Produkt. Die fettgedruckten Bezeichnungen stellen die zusammengesetzten Lebensmittel dar.

2.2.3 Angabe von Nährwerten

Die EU-Verordnung Nr. 1169/2011 [L60] regelt auch die Angabe von Nährwerteinformationen auf Produktverpackungen. Demnach ist die Angabe von Brennwert bzw. Energiegehalt, Fett, gesättigten Fettsäuren, Kohlenhydraten, Zucker, Eiweiß und Salz kennzeichnungspflichtig. Weitere Angaben, beispielsweise für Vitamine oder Spurenelemente, sind freiwillig [L43].

In den Community-Datenquellen WikiFood.eu, das-ist-drin.de und OpenFoodFacts.org sind viele der Nährwerteinformationen unvollständig. Die am meisten vorhandenen Informationen der Nährstoffe sind die Angaben der Makronährstoffe für Kohlenhydrate, Fett und Proteine. Diese Daten sind im Regelfall neben anderen wie Energiegehalt, gesättigte Fettsäuren, Zucker und Salz auf den Produktverpackungen aufgedruckt. Abbildung 2.3 zeigt als Beispiel die Nährwerteliste des Produktes Steinofen-Pizza der Marke Wagner.

Nährwerte		pro 100 g	pro 160 g (1/2 Pizza)	pro Portion 320 g (1 Pizza)	% Referenz-menge* pro Portion
Energie	kJ	1064	1702	3404	41%
	kcal	254	406	812	
Fett		10,8 g	17,3 g	34,6 g	49%
- davon gesättigte Fettsäuren		4,5 g	7,1 g	14,2 g	71%
Kohlenhydrate		27,0 g	43,2 g	86,4 g	33%
- davon Zucker		3,9 g	6,2 g	12,4 g	14%
Ballaststoffe		2,1 g	3,3 g	6,6 g	–
Eiweiß		11,1 g	17,7 g	35,4 g	71%
Salz		1,3 g	2,1 g	4,1 g	68%

*Referenzmenge für einen durchschnittlichen Erwachsenen (8400 kJ/2000 kcal). Packung enthält 1 Portion.

Abbildung 2.3 Beispiel einer Nährwerteliste des Produktes Steinofen-Pizza der Marke Wagner

Nach Biesalski et al. (2020) sind Makronährstoffe die Grundbausteine von Lebensmitteln, die dem menschlichen Körper Energie liefern [12]. Mikronährstoffangaben enthalten Informationen über Mineralstoffe und Spurenelemente (wie z. B. Magnesium, Calcium, Kalium, Zink, Kupfer, Eisen, Selen und Jod) und Informationen über Vitamine (z. B. Vitamin A, B1, B2, B6, C, D, E und Beta-Carotin). Diese Informationen sind nur teilweise in den Community-Datenquellen enthalten. Mikronährstoffe sind wichtig für die Aufrechterhaltung des Stoffwechsels, das Wachstum und die Energieproduktion [L30].

2.2.4 Angabe für häufig auftretende Allergene

14 Lebensmittelzutaten und deren Erzeugnisse sind bekannt dafür, dass sie besonders häufig Auslöser von Nahrungsmittelunverträglichkeiten oder -allergien sind [59]. Nachfolgend sind 14 Auslöser aufgelistet.

- Glutenhaltiges Getreide
- Krebstiere
- Eier
- Fische
- Erdnüsse
- Sojabohnen
- Milch, Laktose
- Schalenfrüchte
- Sellerie
- Senf
- Sesamsamen
- Schwefeldioxid und Sulfite
- Lupinen
- Weichtiere

Diese 14 Auslöser müssen nach EU-Verordnung Nr. 1169/2011 [L60] speziell auf den Produktverpackungen gekennzeichnet werden [L6]. Auf manchen Produktverpackungen werden diese fettgedruckt in den Inhaltsstofflisten dargestellt.

2.3 Verwendung von Lebensmittel- und Studiendaten aus dem Projekt „Digitale Dienstleistungen in der Ernährungsberatung“ (DiDiER)

Das vom BMBF geförderten Projekt DiDiER beschäftigte sich in den Jahren 2016 bis 2019 mit der Verbesserung von Dienstleistungen in der Ernährungsberatung, speziell in den beiden Bereichen Gebrechlichkeit durch Mangelernährung (Frailty) und Lebensmittelunverträglichkeiten und -allergien. Innerhalb des Projektes wurden digitale Lösungen entwickelt, die Beratungsprozesse im Gesundheitswesen verbessern sollen [26]. Während des DiDiER-Projektes wurden

Lebensmitteldatensätze aus verschiedenen Datenquellen genutzt, die in der vorliegenden Arbeit Anwendung finden. Des Weiteren wurden Teile einer bestehenden Lebensmitteldaten-API wiederverwendet bzw. optimiert, mit Big Data Technologie gekoppelt und weiterentwickelt. Während der DiDiER-Probandenstudie wurden elektronische Ernährungstagebücher (mit Bezug zu Nahrungsmittelunverträglichkeiten und -allergien) ausgewertet, um Zusammenhänge zwischen Lebensmitteldaten und Symptomen zu erkennen. In diesem Zuge wurden Lebensmitteldaten gesammelt, die bei der Evaluierung der, während dieser Arbeit entwickelten Ähnlichkeits- und Plausibilitätsanalyse verwendet werden konnten [62][64].

2.4 Zur Verfügung stehende Datenquellen

Als Datenlieferant für das FDWH dienen verschiedene Arten von Datenquellen und Datenbanken. Dabei wird zwischen sogenannten Datenbanken für Lebensmittelzusammensetzungen (FCDB, auf Englisch: Food Composition Database) und Lebensmittelproduktdatenbanken (FPDB, auf Englisch: Food Product Database) unterschieden. Die FCDB hat zum größten Teil nur die Daten von natürlichen unverpackten Lebensmitteln oder von ganzen aus den Lebensmitteln zusammengesetzten Gerichten gespeichert. Diese Daten enthalten unter anderem Informationen über den Namen des Lebensmittels, Verarbeitungsformen und seine Bestandteile wie z. B. Makronährstoffe, Vitamine und Spurenelemente [L34]. Die FPDB besteht überwiegend aus Datensätzen von verpackten Lebensmittelprodukten. Diese Daten enthalten unter anderem die im Folgenden aufgelisteten Informationen (nachfolgend auch als Datenattribute bezeichnet):

- Produktname
- Markenname
- Herkunftsland
- GTIN
- Inhaltsmenge
- Inhaltsstoffliste
- Nährwerteinformationen
- Allergieinformationen

Die FCDB-Daten helfen im weiteren Verlauf der Arbeit sowohl durch das Liefern wichtiger Informationen von natürlichen unverpackten Lebensmitteln (wie z. B. Apfel, Brot oder Fleisch) als auch durch dessen Verwendung der Qualitätsverbesserung der FPDB-Daten, indem sie wichtige Informationen über Inhaltsstoffe dieser Daten liefern. Rund 15 000 Daten werden vom Max Rubner-

Institut (MRI), dem Forschungsinstitut für Ernährung und Lebensmittel, unter dem Namen Bundeslebensmittelschlüssel (BLS) zur Verfügung gestellt [L35].

Nachfolgend folgt eine Beschreibung der im FDWH verwendeten Datenquellen. Ausschließlich die BLS-Datenbank ist eine FCDB (Food Composition Database). Alle anderen Datenquellen sind FPDB (Lebensmittelproduktdatenbanken).

2.4.1 Bundeslebensmittelschlüssel (BLS)

Entwickelt wurde die BLS-Datenbank, um die Energie- und Nährstoffzufuhr von ernährungsepidemiologischen Studien und Erhebungen des Lebensmittelverzehrs zu analysieren und auszuwerten [L35]. Da der BLS-Datenbestand vom MRI und deren Partnern kontinuierlich geprüft, überarbeitet und optimiert wird, ist er eine seriöse Datenquelle für Lebensmittelinformationen mit vertrauenswürdigen Daten [L35]. Der BLS enthält Daten über herkömmliche unverpackte Lebensmittel und zusammengesetzten Gerichten, jedoch keine Daten über Markenprodukte. Deshalb enthält er auch keine Informationen zu Markennamen, Vertriebsland oder eine damit verbundene GTIN. Zudem fehlen hier Inhaltsstofflisten mit weiteren enthaltenen Lebensmittelproduktbestandteilen (z. B. Wasser, Zucker etc.), kennzeichnungspflichtige Allergene oder Verpackungsinformationen. Jeder Datensatz der BLS-Datenbank repräsentiert ein Lebensmittel oder ein aus mehreren Lebensmitteln zusammengesetztes Gericht (z. B. Spaghetti Bolognese). Die Datenbank beinhaltet als Informationen nur den BLS-Datensatzschlüssel (SBLS), den Namen des Lebensmittels (bzw. des Gerichts) und die verschiedenen Nährwerteinformationen, die in den Lebensmitteln enthalten sein können. Im SBLS sind weitere Informationen enthalten, z. B. Informationen über Typ oder Verarbeitungsform des Lebensmittels, die durch sieben Stellen codiert sind. Tabelle 2.2 enthält eine detaillierte Beschreibung der Bedeutung von jeder der einzelnen Stellen im SBLS [L34].

Stelle im BLS	**Codierung**	**Beschreibung**	**Beispiel**
1. Stelle	Buchstaben von B bis Y	Lebensmittel-Hauptgruppe (Art des Lebensmittels). X und Y stehen für Menükomponenten aus Haushalt und Gastronomie.	B für „Brot und Kleingebäck“

2. Stelle	Zahlen von 0 bis 9	Lebensmittel-Untergruppe.	B1 für „Vollkornbrote“
3. und 4. Stelle	Zahlen von 0 bis 9	fortlaufende, klassifizierende Verschlüsselung der einzelnen Lebensmittel	B111 für „Vollkornbrot-Weizenvollkornbrot“
5. Stelle	Zahlen von 0 bis 9	mehrfache Bedeutung	industrielle Verarbeitung, Fettgehalt bei Milchprodukten, Getränke mit Milch/Zucker/Alkohol etc., weitere Spezifizierung der Einzellebensmittel.
6. Stelle	Zahlen von 0 bis 9	Garverfahren (bei den Hauptgruppen B bis W), Zubereitungsinformationen (bei den Hauptgruppen X und Y)	3 für „gekocht“.
7. Stelle	Zahlen von 0 bis 9	Gewichtsbezug	0, 2 und 3 sind Verzehrgewicht ohne Küchenabfall, 1 und 4 sind Verzehrgewicht mit Küchenabfall

Tabelle 2.2: Beschreibung der Stellen im SBLS

Der BLS-Schlüssel befindet sich in einer hierarchischen Struktur. Jede der Stellen steht für den Gruppenbegriff der jeweils nächsten Stellen. Wenn eine Stelle gestrichen wird, wird sie auf 0 gesetzt. Die nachfolgenden Stellen stehen dann für den weiteren Gruppenbegriff [L34].

Interessant für die Analysen in Gesundheitsapps sind auch die vom BLS bereitgestellten Verarbeitungsformen und -prozesse wie beispielsweise roh, gekocht oder gesalzen.

Im Folgenden sind die Lebensmittel-Hauptgruppen des BLS aufgelistet, an denen sich zum Teil auch die Kategorien des FDWH orientieren (Daten vom Max Rubner-Institut [L34]).

- B: Brot und Kleingebäck
- C: Cerealien, Getreide, Reis
- D: Backwaren
- E: Eier und Teigwaren
- F: Früchte und Obst
- G: Gemüse
- H: Hülsenfrüchte, Schalenobst und Samen
- K: Kartoffeln, stärkereiche Pflanzenteile, Pilze
- M: Milch und Käse
- N: Nichtalkoholische Getränke
- P: Alkoholische Getränke
- Q: Öle, Fette, Butter, Schmalz
- R: Gewürze, Würzmittel, Hilfsstoffe
- S: Süßwaren, Zucker, Schokolade, Eis
- T: Fisch und Meeresfrüchte
- U: Rind-, Kalb-, Schweine-, Hammel- und Lammfleisch
- V: Wild, Geflügel, Federwild, Innereien
- W: Wurst und Fleischwaren
- X: überwiegend pflanzliche Menükomponenten
- Y: vorwiegend tierische Menükomponenten

Zum besseren Verständnis des SBLS dient das folgende Beispiel. Dem Lebensmittel „Ravioli mit Fleischfüllung gekocht“ wird der Schlüssel E603032 zugeordnet. Die erste Stelle besagt, dass das Lebensmittel in die Hauptgruppe „Eier und Teigwaren“ eingeordnet ist. Die zweite Stelle bezieht sich auf die Untergruppe „Teigwaren besonderer Art“. Die Stellen 3 und 4 beschreiben das Einzellebensmittel „Ravioli mit Fleischfüllung“. Stelle 5 ist die Zahl 0 und beinhaltet somit keine Information. Stelle 6 bezieht sich auf das Garverfahren und besagt, dass das Lebensmittel gekocht wurde. Stelle 7 beschreibt, dass sich das Gewicht der einzelnen Nährstoffe im Datensatz auf das Lebensmittel ohne Küchenabfall im abgegossenen Zustand bezieht [L34].

Die BLS-Datenbank enthält 139 verschiedene Nährwerte-Attribute (welche verschiedene Arten von Mikro- und Makronährstoffen repräsentieren) die dem jeweiligen SBLS zugeordnet sind. Diese Attribute enthalten somit Informationen über die Zusammensetzung der Lebensmittel. Darunter sind Makronährstoffe (Energie-, Fett-, Protein- und Kohlenhydrate-Gehalt etc.) und Mikronährstoffe wie

Vitamine (Vitamin A, B1, B2, B3, C, D, E etc.) sowie Mineralstoffe (Natrium, Kalium, Calcium, Magnesium etc.). Weiterhin enthalten die Nährwerte-Attribute Informationen über Spurenelemente, Kohlenhydratzusammensetzungen, Ballaststoffzusammensetzungen, Aminosäuren, Fettzusammensetzungen und Gesamtkennzahlen [L34].

2.4.2 WikiFood.eu

WikiFood wurde 2006 vom CRP Henri Tudor in Luxemburg (seit 2015 LIST) ins Leben gerufen. Die Plattform stellt eine Webseite zur Suche und Pflege von Lebensmittelproduktdaten zur Verfügung. Die Daten von WikiFood stammen von Nutzern, der Industrie, Herstellern und anderen Datenanbietern. WikiFood beinhaltet das Wiki-Prinzip, das auf der Idee der Online-Enzyklopädie Wikipedia basiert. Durch dieses werden mittels Crowd-Sourcing durch Teilnehmer der WikiFood-Community Daten mit Informationen über Lebensmittelprodukte zusammengetragen, überprüft, korrigiert und verwaltet [60]. WikiFood hat seine Daten für die Forschung dem Konsortium des Projektes “Digitale Dienstleistungen in der Ernährungsberatung” (DiDiER) über eine XML-API (API, deren Anfrage und Antwort durch die menschen- und maschinenlesbare Formatierungssprache XML beschrieben wird [8][32]) sowie als Datenbank-Dump zur Verfügung gestellt. Die Plattform von WikiFood hält Datensätze von Lebensmittelprodukten überwiegend aus Deutschland, Frankreich, Belgien und Luxemburg bereit. Die Webseite von WikiFood wurd 2019 vom Netz genommen. Nachfolgend sind die Informationen (bzw. Daten-Attribute) aufgelistet, die von der WikiFood-Datenbank stammen in das FDWH integriert werden.

- EAN-Barcode
- Produktname
- Markenname
- Inhaltsmenge
- Kategorien (von den Entwicklern der WikiFood-Plattform ausgewählt)
- Herkunftsland
- Inhaltsstoffliste
- Nährwerte (Makronährstoffe, Mikronährstoffe)
- Angaben über kennzeichnungspflichtige Allergene in den Produkten
- Zusätzliche auf Produktverpackung aufgedruckte Informationen

Abbildung 2.4 stellt einen Auszug aus dem Entitätsbeziehungsmodell (Entity Relationship Modell, ERM) der WikiFood-Datenbank grafisch dar. Im weiteren Verlauf der Arbeit gilt es, die Informationen

in Bezug auf die im ERM dargestellten Daten-Attribute zu extrahieren und vereinheitlicht mit den Daten aus den anderen Quellen im FDWH zu speichern.

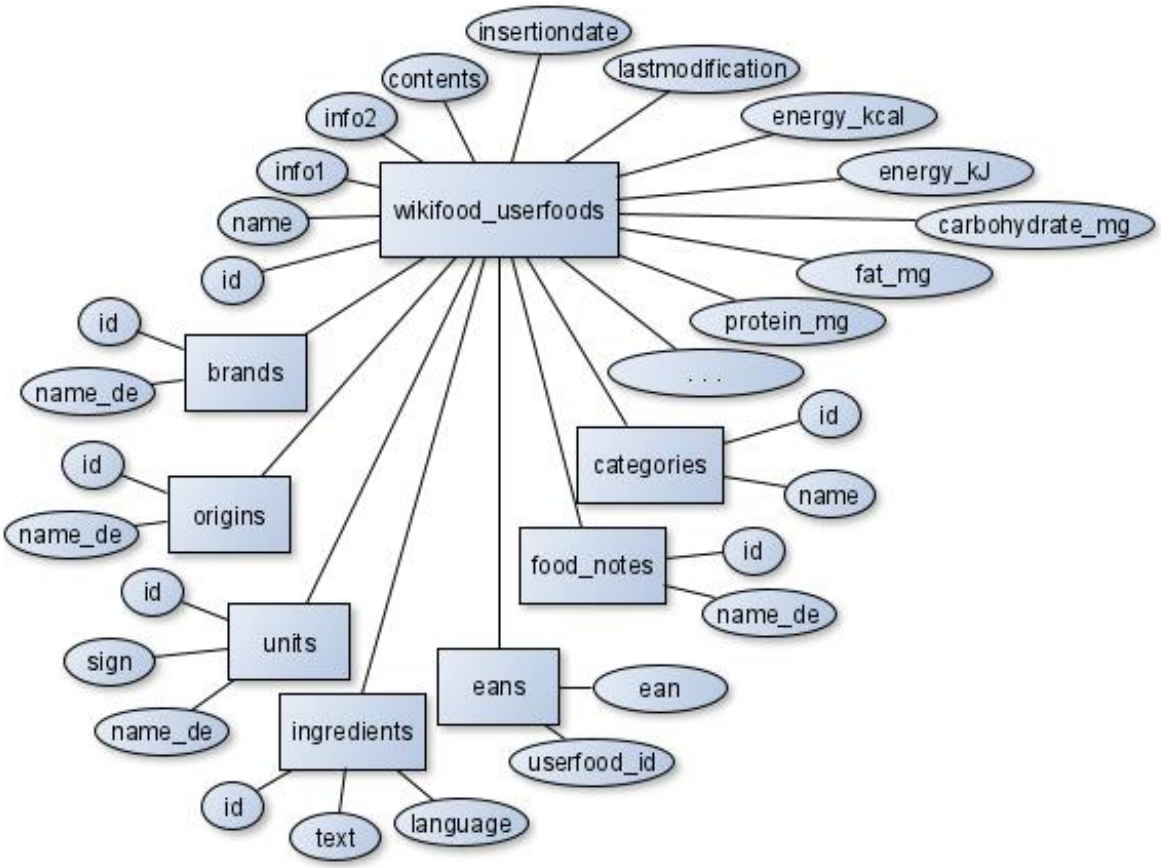


Abbildung 2.4 Auszug aus ERM der WikiFood-Datenbank

2.4.3 Danone

Die Danone GmbH ist ein Hersteller von Milchprodukten [L12]. Auch Danone hat Daten seiner Lebensmittel für Forschungszwecke zur Verfügung gestellt. Die folgende Auflistung beinhaltet die Informationen, die von der Danone GmbH zur Verfügung gestellt wurden.

- Produktname
- Inhaltsmenge
- Inhaltsstoffliste
- Nährwerte (Makronährstoffe, Mikronährstoffe)

- Angaben über kennzeichnungspflichtige Allergene in den Produkten
- Zusätzliche auf Produktverpackung aufgedruckte Informationen

2.4.4 OpenFoodFacts.org

OpenFoodFacts.org ist eine Web-Plattform, die Lebensmittelproduktdatensätze sowohl über eine API als auch über einen sogenannten Datenbank-Dump (eine Kopie aller Daten der kompletten Datenbank) zur Verfügung stellt. Die Datenbank-Dumps stehen im CSV-Format (Datentabellen in Kommagetrennter Darstellung [49]) und im MongoDB-Format (sogenannte Not-only-SQL Datenbank, abgekürzt: NoSQL mit Daten im JSON-Format [16]) zur Verfügung. Die OpenFoodFacts-Plattform wurde von Stéphane Gigandet in Frankreich gegründet, der als Betreiber mehrerer Food-Blog-Portale aktiv ist. Das OpenFoodFact-Projekt wird durch eine große internationale Internet-Community aus lebensmittelinteressierten Menschen unterstützt, welche die Einträge der Lebensmittelproduktdatenbank regelmäßig überarbeiten und neue Produkte hinzufügen [L52]. Die bereitgestellten Datensätze sind weltweit für alle Anwender frei zugänglich und unterliegen der Open Database Licence (ODbL Version 1.0, die Daten dürfen frei verwendet, verändert und freigegeben werden [L41]). Sie beinhalten des Weiteren Daten aus mehreren Ländern (darunter auch Deutschland) und stehen in mehreren Sprachen zur Verfügung (unter anderem in der deutschen, englischen und französischen Sprache). OpenFoodFacts.org beinhaltet zusätzlich einen Blog, über den sich die Community über Lebensmittelinformationen austauscht und diese diskutiert [L52].

Zusätzlich stellt die Plattform eine App zur Verfügung, innerhalb derer sich Informationen über Lebensmittelprodukte anhand des Scannens ihres GTIN-Barcodes und über die Eingabe eines Suchwortes in eine Suchmaske ermitteln lassen. Auch über die Webseite OpenFoodFacts.org lassen sich Produktinformationen über eine Suchmaske abfragen. Die Datensätze der OpenFoodFacts.org-Plattform [L52] beinhalten unter anderem die im Folgenden gelisteten Informationen.

- EAN-Barcode
- Produktname
- Markenname
- Inhaltsmenge
- Kategorie-Label (durch den Autor des jeweiligen Datensatzes festgelegt)
- Herkunftsland, Verkaufsland
- Inhaltsstoffliste
- Nährwerte (Makronährstoffe, Mikronährstoffe)
- Angaben über kennzeichnungspflichtige Allergene in den Produkten

- Zusätzliche auf Produktverpackung aufgedruckte Informationen

Abbildung 2.5 zeigt einen Auszug aus der Liste der Attributfelder, in denen die Daten von OpenFoodFacts.org in der MongoDB gespeichert sind [L52].

```
List of fields:

# general information:

code : barcode of the product (can be EAN-13 or internal codes for some food stores),
       for products without a barcode, Open Food Facts assigns a number starting with the 200 reserved prefix
url : url of the product page on Open Food Facts
created_datetime : date that the product was added (iso8601 format: yyyy-mm-ddThh:mn:ssZ)
last_modified_datetime : date that the product page was last modified (iso8601 format: yyyy-mm-ddThh:mn:ssZ)
product_name : name of the product
quantity : quantity and unit

# tags:

packaging : shape, material
brands
categories_tags
origins_tags
manufacturing_places_tags
labels
countries : list of countries where the product is sold

# ingredients:

ingredients_text
traces_tags

# nutrition facts:

energy-kj_100g
energy-kcal_100g
proteins_100g
carbohydrates_100g
fat_100g
...
```

Abbildung 2.5 MongoDB von OpenFoodFacts.org: Auszug der Attributfelder

2.4.5 FoodRepo.org

FoodRepo.org ist wie OpenFoodFacts.org eine freie internationale Web-Plattform, die kostenlose Daten mit Informationen über Lebensmittelprodukte aus den Ländern Deutschland, Frankreich, Italien und der Schweiz liefert. Die Plattform wurde durch das Digital Epidemiology Lab (EPFL) gegründet [L53]. Die Daten werden durch eine Internet-Community mit derzeit über 2200 Nutzern (Stand 22. Oktober 2021) bereitgestellt und bearbeitet. Über eine API können die Daten im menschen- und maschinenlesbaren JSON-Format [14] abgerufen werden. Auch eine Funktion für einen Download aller Daten, die FoodRepo.org bereithält, ist möglich (auch im JSON-Format). Die Ergebnisse der Produktabfrage werden sowohl in der deutschen als auch in der englischen, französischen und italienischen Sprache

geliefert. Folgende Informationen werden durch die Daten der FoodRepo.org-Plattform bereitgestellt [L53].

- EAN-Barcode
- Produktname inklusive des Markennamens
- Inhaltsmenge
- Herkunftsland
- Inhaltsstoffliste
- Nährwerte (Makronährstoffe)
- Zusätzliche auf Produktverpackung aufgedruckte Informationen

2.4.6 das-ist-drin.de

Das-ist-drin.de ist ähnlich wie WikiFood.eu eine Plattform, in der jeder Anwender Lebensmittelproduktinformationen nach dem Wiki-Prinzip einfügen und bearbeiten kann. In einem Webportal lassen sich die Produkte über eine Suchmaske abfragen [L46]. das-ist-drin.de bietet zusätzlich eine App an, welche die Daten über eine EAN-Barcode-Scanfunktion und über eine Suchmaske liefert. Die das-ist-drin.de-Plattform wurde von der Firma snoopmedia GmbH gegründet, welche dem DiDiER-Konsortium 1000 Datensätze von in Deutschland erhältlichen Produkten für Forschungszwecke zur Verfügung gestellt hat. Die das-ist-drin.de-Plattform liefert die nachfolgend aufgelisteten Informationen [L46].

- EAN-Barcode
- Produktname
- Hersteller inklusive Markenname
- Inhaltsstoffliste
- Nährwerte (Makronährstoffe, Mikronährstoffe)
- Zusätzliche auf Produktverpackung aufgedruckte Informationen

Abbildung 2.6 stellt einen Auszug einer Produktdatenselektierung auf der das-ist-drin.de-Webseite grafisch dar.

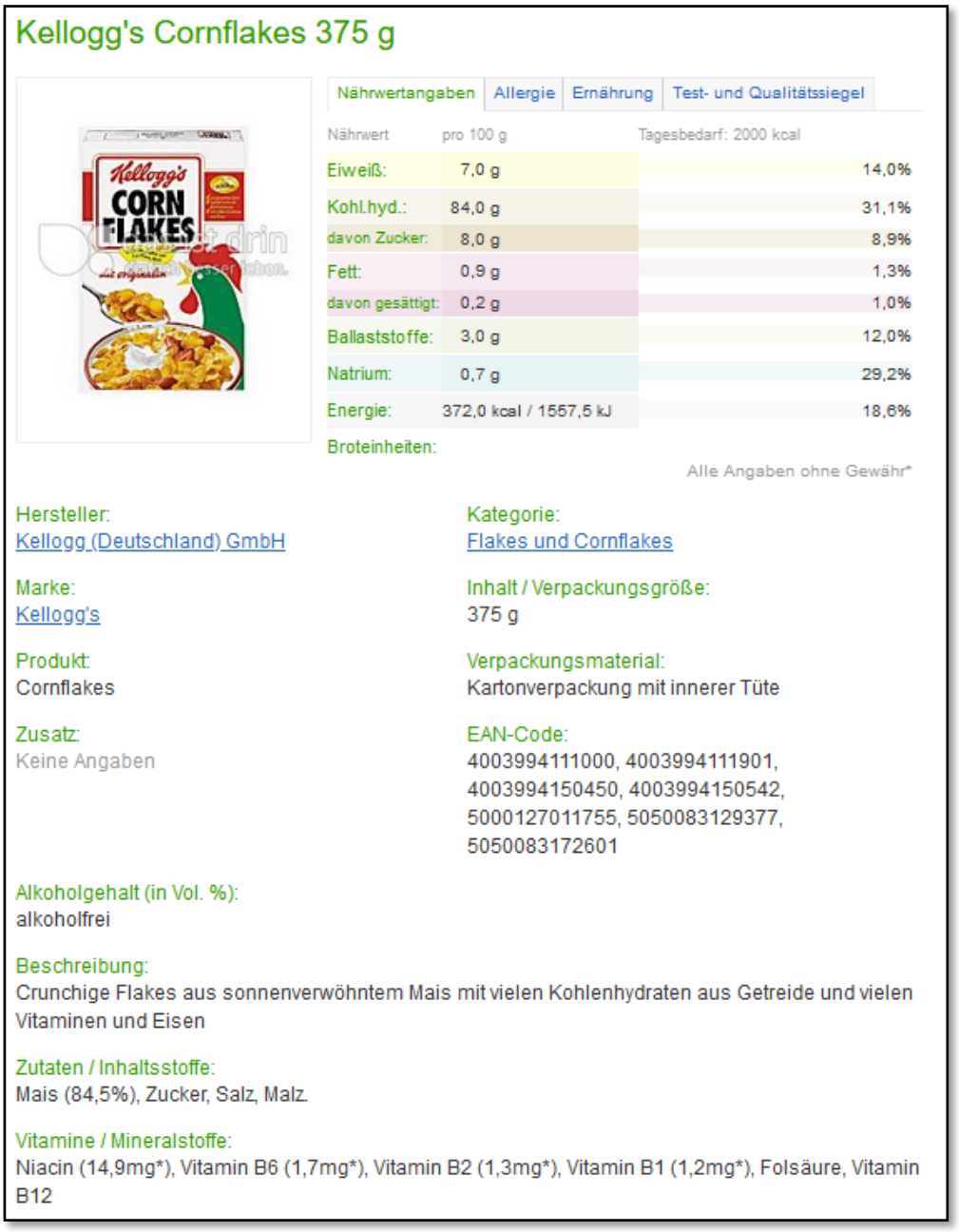


Abbildung 2.6 Auszug einer Produktdatenselektierung auf der das-ist-drin.de Webseite

2.5 Attribute und Datenmerkmale der Lebensmitteldatenquellen

Unter dem Begriff Attribut werden die jeweiligen Bezeichnungen der Produktmerkmale zusammengefasst. Im FDWH werden die Attribute meist durch ein oder mehrere Datenfelder (z. B. Wert und Einheit für die Inhaltsmenge) repräsentiert. Somit enthält beispielsweise ein Produkt die Attribute Produktname, Inhaltsstoffliste und jeweils Attribute für die enthaltenen Nährstoffe (z. B. Kalorien, Fettgehalt, Proteine, Kohlenhydrate, etc.). Solche Attribute kann man auch mit der Spalte einer

relationalen Datenbank-Tabelle vergleichen. Tabelle 2.3 zeigt die für das FDWH relevanten Attribute, deren Beschreibung und bietet eine Übersicht welche Datenquelle die jeweiligen Informationen liefert. Aufgelistet sind nur Attribute die potenziell für Gesundheitsapps mit dem Bezug zur Ernährungsmedizin von Bedeutung sind.

Attribut (Bezeichnung im FDWH)	**Beschreibung**	**Verknüpfung mit Daten aus Datenquelle**
Lebensmittel-/Produkt-Name (food_name)	Bezeichnung des Lebensmittels/des Produktes.	Bundeslebensmittelschlüssel, WikiFood.eu, Danone, OpenFoodFacts.org, FoodRepo.org, das-ist-drin.de
Markenname (brand_name)	Bezeichnung der Marke des Produktes.	WikiFood.eu, OpenFoodFacts.org, FoodRepo.org, das-ist-drin.de
Herkunftsland (origin_name)	Herstellungs-/Herkunftsland des Produktes.	WikiFood.eu, OpenFoodFacts.org, FoodRepo.org
GTIN/EAN (gtin_number)	Global Trade Item Number bzw. European Article Number.	WikiFood.eu, OpenFoodFacts.org, FoodRepo.org, das-ist-drin.de
Inhaltsmenge (content_value)	Angabe über Menge des Inhalts der Produktverpackung.	WikiFood.eu, Danone, OpenFoodFacts.org, FoodRepo.org
Inhaltsstoffliste (ingredient_list)	Liste mit Zutaten aus denen das Produkt zusammengesetzt ist.	WikiFood.eu, Danone, OpenFoodFacts.org, FoodRepo.org, das-ist-drin.de
Nährwerteinformationen (Attribute für jeden der gespeicherten Nährwerte: energy_kcal, carbohydrates, fat, proteins, etc.)	Angabe über Menge und Art der Mikro- und Makronährstoffe die im Produkt enthalten sind.	Bundeslebensmittelschlüssel, WikiFood.eu, Danone, OpenFoodFacts.org, FoodRepo.org, das-ist-drin.de

Häufig auftretende Allergene	Angabe über Allergene, die in den Produkten enthalten und kennzeichnungspflichtig sind.	WikiFood.eu, Danone, OpenFoodFacts.org

Tabelle 2.3 Für das FDWH relevante Attribute mit Bezug zu den Datenquellen, in denen sie enthalten sind

Die Attribute für Lebensmittel- und Produktname, Markenname, Herkunftsland, GTIN bzw. EAN und Inhaltsmenge dienen zur Identifikation des jeweiligen Produktes. Inhaltsstoffliste, Nährwerteinformationen und häufig auftretende Allergien beschreiben die Zusammensetzung des jeweiligen Lebensmittels oder Lebensmittelproduktes.

Drei der insgesamt sechs genannten externen Datenquellen haben Informationen über diese Hauptallergene in unterschiedlich strukturierten Datenfeldern aufgeführt. WikiFood.eu gibt in einer durch Kommas separierten Zeichenkette an, welche Allergene sicher nicht enthalten sind, beispielsweise „Nussfrei, Glutenfrei“. OpenFoodFacts.org liefert eine Zeichenkette, die Allergen-Labels beinhaltet, die darüber informieren, welche Allergene und Spuren von Allergenen enthalten sind (z. B. „en:nuts, en:gluten“). Der Datensatz von Danone enthält sechs Datenfelder für die Auslöser Gluten, Laktose, Soja, Nüsse, Erdnüsse und Ei. Jedes der Felder enthält entweder das Zeichen „j“, wenn der Auslöser im Produkt enthalten ist oder „n“, wenn er nicht enthalten ist. Im späteren Transforming-Prozess (siehe Abschnitt 3.3.3) müssen die Informationen aus den unterschiedlich strukturierten Datenfeldern extrahiert und in das einheitliche Datenformat des FDWH transformiert werden.

Weitere nicht in Tabelle 2.3 enthaltenen Informationen sind sonstige Informationen über Verpackungsart oder Beschreibungen des Lebensmittels. Diese Informationen werden im FDWH unter dem Attributfeld „additional_information“ (zusätzliche Informationen) als Zeichenkette festgehalten. Die Informationen sind nicht für die in Abschnitt 2.1 beschriebenen Anwendungszwecke relevant, können aber für den Anwender hilfreiche Informationen zur Identifikation eines Produktes liefern, für das Informationen aus dem FDWH selektiert werden sollen. Informationen über GTIN, Inhaltsstoffliste, Nährwerte und häufig auftretende Allergene, die teilweise auch auf Produktverpackungen aufgedruckt sind bzw. aufgedruckt sein müssen, werden im Anschluss näher erläutert.

2.6 Meta- und Wissensdaten

Für die verschiedenen Prozesse zur Extrahierung, Transformierung und dem Laden der Lebensmitteldaten ins FDWH werden zusätzliche Daten, sogenannte Metadaten benötigt.

Als Metadaten werden Informationen über Eigenschaften und Bedeutung der Daten (sogenannte Metainformationen) von Datenbank- und Data-Warehouse-Systemen genannt [9][21]. In dieser Arbeit werden vorliegende Informationen der externen Datenquellen und des FDWH als Metadaten gespeichert. Dazu gehören auch Schemadaten der Datenbank, in welcher die Daten des FDWH gespeichert sind. Die Metadaten sind, wie in Abschnitt 3.2.3 beschrieben, in spezielle INI-Dateien integriert (Dateien mit der Dateiendung .ini zur initialen Speicherung von Konfigurationsdaten), die Metainformationen zur Konfiguration des Transformationsprozesses zur einheitlichen Datenspeicherung enthalten. Metadaten der externen Lebensmitteldatenquellen werden während des Data Profiling Prozesses ermittelt und dienen der Analyse von Datentypen und Regeln zur Datentransformation.

Wie Castillo & Jorzyk beschreiben, werden Informationen in Daten zu Wissen, wenn sie für den Empfänger der Information einen nützlichen und neuen Sachverhalt liefern [L9]. Dies umfasst auch externe Informationen mit Wissen, welches im FDWH beispielsweise zur Vereinheitlichung von Datenattributwerten nützlich ist oder Informationen die Wissen über gespeicherte Daten beinhalten. Somit können Metadaten auch als Wissensdaten bezeichnet werden.

Im FDWH sind zudem Wissensdaten gespeichert die außer den technischen Eigenschaften eines Attributes, z. B. den Datentyp bzw. das Datenformat und die Datenstruktur, weiteres Wissen enthalten (beispielsweise standardisierte Bezeichnungen oder Synonyme zu den als Zeichenkette gespeicherten Namen oder Bezeichnern). Zu solchen Daten gehören ebenso die Informationen der standardisierten Liste von SI-Einheiten nach ISO 80000-1:2009 [L24]), die den Mengenangaben von Inhalt und verschiedener Nährwerte dienen. Des Weiteren werden die Daten der Liste mit standardisierten Ländercodes nach ISO 3166 [L25] zur Identifizierung der Herkunft der Lebensmittel verwendet. Außerdem werden die Daten der Liste mit Sprachcodes nach ISO 639-1:2002 [L23] für die Angabe der jeweiligen Sprache in Inhaltsstofflisten verwendet. Das jeweilige Herkunftsland eines Datensatzes kann anhand seiner gespeicherten Adresse nachvollzogen werden. Dazu wird eine Verknüpfung zu der GeoLite2 City Database, eine Datenbank der Firma MaxMind Inc. eingesetzt, die Straßennamen, Postleitzahlen, Städte- und Dörfer-Namen verknüpft mit den jeweiligen Länderbezeichnungen enthält [L36][A].

Die in der BLS-Datenbank gespeicherten Angaben von Mikro- und Makronährstoffen der enthaltenen Lebensmittel sind weitestgehend vom MRI recherchiert bzw. analysiert und korrekt eingetragen [L35][L28]. Aufgrund der Vertrauenswürdigkeit dieser Daten besteht die Möglichkeit durch eine

Verknüpfung die fehlenden Nährwerteinformationen, die in den Produktdatensätzen der FPDB fehlen, über ihre oder vergleichbare ähnliche Inhaltsstofflisten zu ermitteln. Ein großer Teil aller Inhaltsstoffe, die in den Inhaltsstofflisten der FPDB genutzt werden, sind als BLS-Datensätze gelistet. Die BLS-Datenbank fungiert somit, außer als Datenlieferant für die API des FDWH, auch als Wissensdatenbank mit den Informationen über den Gehalt bestimmter Nährwerte in den Lebensmitteln. Hierzu ist es notwendig den prozentualen Anteil eines Inhaltsstoffes unter allen Zutaten in der Inhaltsstoffliste eines jeweiligen Lebensmittelproduktes zu kennen. Dieser Anteil ist in vielen der Inhaltsstofflisten in der FPDB angegeben, aber nicht in allen.

2.7 Datenqualität

Nach Delone & McLean (1992) gibt es eine enge Verbindung zwischen Datenqualität und Anwenderzufriedenheit [20]. Ist die Qualität der Daten einer Anwendung unzureichend, wird die Anwendung nur geringe Akzeptanz bei potenziellen Anwendern finden. Im schlimmsten Fall wird diese für den Zweck der Anwendung nutzlos. In diesem Abschnitt wird erläutert, welche Kategorien und Arten von Qualität für Lebensmitteldaten beachtet werden müssen, damit diese für die vorhergesehenen Zwecke anwendbar sind.

Wang und Strong (1996) fanden in einer Studie über Datenqualitätsmerkmale heraus, dass Genauigkeit und Korrektheit als wichtigste Qualitätsmerkmale bei Endanwendern zählen [73]. Außerdem wurden weitere Qualitätsmerkmale in vier Kategorien eingeteilt werden, welche nachfolgend aufgelistet sind [73][35].

1. Kategorie „Innere Datenqualität“:
    - Glaubwürdigkeit, Genauigkeit, Objektivität, Vertrauenswürdigkeit
2. Kategorie „Kontextabhängige Datenqualität“:
    - Zusatznutzen, Relevanz, Aktualität, Vollständigkeit, angemessenes Datenvolumen
3. Kategorie „Darstellungsqualität“:
    - Interpretierbarkeit, Verständlichkeit, konsistente Darstellung, knappe Darstellung
4. Kategorie „Zugangsqualität“:
    - Zugriffsmöglichkeit, Zugriffssicherheit

Nach der Norm DIN EN ISO 9001:2015 [L13] beschreibt Datenqualität die Korrektheit, die Relevanz und die Verlässlichkeit von Daten abhängig vom Zweck, den die Daten erfüllen sollen. Nachfolgend gelistete Qualitätsmerkmale werden hierbei ebenfalls in vier Kategorien zusammengefasst.

1. Inhalt
    - Hohes Ansehen, Fehlerfreiheit, Objektivität, Glaubwürdigkeit
2. Nutzung
    - Aktualität, Wertschöpfung, Vollständigkeit, Relevanz
3. Darstellung
    - Übersichtlichkeit, Verständlichkeit, einheitliche Darstellung, eindeutige Auslegbarkeit
4. System
    - Zugänglichkeit, Bearbeitbarkeit

Die beiden Auflistungen nach Wang und Strong (1996) und der Norm DIN EN ISO 9001:2015 sagen in den meisten Punkten das Gleiche aus [73]. Im Bereich der inneren Datenqualität bzw. des Inhaltes sollen die Daten glaubwürdig und demnach auch vertrauenswürdig, objektiv, genau und fehlerfrei sein. Die kontextabhängige Datenqualität bzw. der damit verbundene Nutzen beschreibt die Aktualität, Vollständigkeit, Relevanz, Wertschöpfung und angemessenes Datenvolumen, wobei Letzteres im aktuellen Zeitalter von Big Data oft vernachlässigbar ist, da Speicherplatz nicht mehr so teuer ist wie zu der Zeit der Veröffentlichung von Wang & Strong (1996) [73].

Die Darstellungsqualität beschreibt eine knappe und einheitliche Darstellung, Übersichtlichkeit und Verständlichkeit, eine knappe und konsistente Darstellung sowie die eindeutige Auslegbarkeit bzw. Interpretierbarkeit von Daten. Die Kategorie System bzw. der Zugang zu diesem erfordert eine bestmögliche Zugriffsmöglichkeit bzw. Zugänglichkeit, Bearbeitbarkeit und Zugriffssicherheit. Die verschiedenen Merkmale bedingen sich zum Teil gegenseitig.

Die Qualität im Bereich Glaubwürdigkeit, Vertrauenswürdigkeit und Wertschöpfung soll durch Plausibilitätsanalysen und dem Vergleich der Datensätze untereinander erzielt werden. Objektive, genaue und relevante Lebensmitteldatensätze werden bestmöglich durch Vereinheitlichung der Daten und der Datenverarbeitungsprozesse sowie der bereits beschriebenen Untersuchungen möglicher Anwendungszwecke und der zur Verfügung stehenden Datenquellen erreicht. Durch die Anwendung und Entwicklung der (zum Teil im nächsten Kapitel beschriebenen) Datenanalyse-Verfahren sollen durch eine entwickelte Lebensmitteldaten- und Lebensmittelproduktdaten-API auf Fehlerfreiheit und Vollständigkeit (ggf. auch auf Aktualität, je nach Alter des eingefügten Datensatzes) geprüfte Daten in Abhängigkeit zum angegebenen Anwendungskontext geliefert werden. Die Daten sollen dabei in einem übersichtlichen und einheitlichen Datenformat in einer konsistenten Darstellung angezeigt werden. Ausführliche Metainformationen und Dokumentation sorgt für eine gute Interpretierbarkeit und Verständlichkeit der Daten [35]. Durch die Anwendung von Big Data Technologie, redundanter Datenhaltung und verteilter Systeme soll eine performante und stabile durch Authentifizierungs- und Autorisierungsverfahren geschützte API für den Zugriff auf die Daten geschaffen werden.

2.8 Methoden der Datenverarbeitung und Datenanalyse

In diesem Kapitel werden die in der vorliegenden Arbeit angewandten Methoden im Bereich der Datenverarbeitung, insbesondere Datenspeicherung und Datentransformation sowie der Datenanalyse, inklusive Machine-Learning-Verfahren, erläutert.

2.8.1 Data Profiling

Für ein Projekt, das sich mit der Erfassung und Verarbeitung von Daten beschäftigt, ist es essenziell ein Verständnis für den Aufbau und die Bedeutung der Daten zu gewinnen. Data Profiling befasst sich mit der Analyse von Daten und deren Metadaten und bringt Methoden mit sich, um Daten auf ihre Struktur, verwendete Datentypen und deren Beschreibung durch Metadaten zu untersuchen [1]. Besonders zur Überführung mehrerer Datenquellen in eine einzige zentrale Datenbank ist es notwendig, Struktur und Bedeutung aller benötigten Datenelemente zu kennen. Die Bedeutung von Datenelementen kann durch den Begriff der Semantik beschrieben werden [22]. Gegenteilig dazu beschreibt die Syntax Regeln, welche die Struktur von Daten festlegen [70]. Dengel & Bernardi (2012) beschreiben, dass die Semantik für den Wissensbegriff und die Wissenspräsentation eine wesentliche Rolle spielen [22]. Zur Beschreibung und Erschließung der Semantik von Daten dienen, wie in Abschnitt 2.6 beschrieben, die Metadaten von Datenbanken [21]. Durch Data Profiling werden sowohl syntaktische Regeln über die Zusammensetzung und Struktur von Daten ermittelt als auch die semantische Bedeutung von Daten [L3].

Im Verlauf der vorliegenden Arbeit werden, nach einer intensiven Analyse von Metadaten der externen Datenquellen, der Bedeutung der Informationen auf Produktverpackungen und der Untersuchung weiter Lebensmittelinformationen, Wissen über die Bedeutung von Lebensmitteldaten zusammengetragen. Aus diesem Wissen wird, das in Abschnitt 3.2.1 dargestellte Ontologie-Modell generiert, das beim Aufbau des FDWH von wichtiger Bedeutung ist. Teile dieses Wissens werden außerdem für Datentransformationsprozesse als Metadaten des FDWH gespeichert.

Durch eine Analyse der Muster- und Datentypen wird ermittelt, ob es sich bei den Daten eines Attributes um numerische Daten oder eine Zeichenkette handelt und welche Art von Zeichen eine solche Zeichenkette beinhaltet. Zudem werden durch Data Profiling fehlende Daten, fehlerhafte Daten oder Dubletten analysiert [43]. Der Data Profiling Prozess muss inkrementell während jedem Transformations-Prozess der Daten durchgeführt werden, um stets in der Lage zu sein, die Inkonsistenzen der neuesten Daten aus externen Quellen zu ermitteln und zu korrigieren. Es werden inkonsistente Datensätze erkannt, die für eine bessere Qualität korrigiert werden müssen und es werden

fehlende Daten oder Dubletten erkannt, die vervollständigt bzw. gelöscht werden müssen [43]. Data Profiling beinhaltet zusammengefasst die folgenden Aufgaben [52][1][A].

- Ermittlung der Bedeutung von Datenattributen und von Beziehungen zwischen Datenattributen eines Datensatzes und spaltenübergreifenden Werteabhängigkeiten in Datentabellen.
  - Damit verbundener Abgleich zwischen Datenattributen zur Inkonsistenz-Erkennung.
- Analyse von Datentypen, Metadaten und Datenmustern zur Regelbildung für die Ermittlung von Spalteneigenschaften zur einheitlichen und konsistenten Speicherung von Datenattributen.
- Verwendung von externen Wissensquellen, um fehlende Werte in den Datensätzen zu berechnen (z. B. Formeln zur Nährwerteberechnung).

2.8.2 Extract, Transform, Load (ETL)

Der ETL-Prozess dient dem Gewinnen benötigter Daten aus den vorliegenden Daten mittels Extraktion, der Standardisierung und Vereinheitlichung der Daten aus den verschiedenen Quellen durch Transformation und dem Laden der transformierten Daten in ein zentrales Data-Warehouse. Bei der Extraktion der Daten aus den verschiedenen Datenquellen müssen die jeweiligen Formate der verschiedenen Daten- und Datenbanktypen berücksichtigt und diesbezüglich muss der Extraktionsprozess angepasst werden [L33][10]. Nach Kimball et al. (1998) umfasst der Transformationsprozess unter anderem die Anpassung von Datentypen, die Konvertierung von Codierungen, Vereinheitlichung von Zeichenketten und Datumsangaben, Umrechnung von Maßeinheiten und die Kombinination bzw. Separierung von Attributwerten [39][10]. Während dieses Prozesses werden mitunter mit den im vorherigen Kapitel vorgestellten Data Profiling Methoden Qualitätsmängel aufgedeckt und wenn möglich bereinigt. Während des Ladeprozesses werden die Datensätze mithilfe der sogenannten Lade-Werkzeuge, die dem Datenbanksystem vorliegen (im Falle der vorliegenden Arbeit mit Hilfe des SQL-Loader der TimescaleDB) im zentralen Data-Warehouse gespeichert. Der Transformationsprozess dient der Vereinheitlichung der Daten. Eine Anwendung der durch Data Profiling generierten Regeln geschieht sowohl während des Transformationsprozesses als auch nach dem Laden der Daten ins zentrale Data-Warehouse als abschließende Qualitätsüberprüfung. Mithilfe von Wissensdaten und Metadaten, die in den Prozessen integriert und in spezifischen Konfigurationsdateien gespeichert sind, werden die Daten der verschiedenen Datenquellen transformiert und im FDWH gespeichert. Die Wissens- und Metadaten enthalten hierbei wichtige Informationen zur Transformation und dienen der abschließenden Qualitätsüberprüfung (siehe Abschnitt 3.3.3).

2.8.3 Data Mining

Nach Cleve und Lämmel (2020) beschreibt der Begriff Data Mining eine Sammlung von Techniken, Methoden und Algorithmen für die Analyse von Daten [19]. Mithilfe von Data Mining wird versucht, bisher unbekannte Informationen aus den Daten zu gewinnen. Data Mining weist große Schnittmengen zum maschinellen Lernen (Machine Learning, ML) auf, wenn es um die angewandten Methoden geht, weshalb die beiden Begriffe oft vermischt werden [19]. Des Weiteren wird in Chamoni (2012) beschrieben, dass Data Mining als eine Herangehensweise analytischer Informationssysteme, die wiederum dem Begriff Business Intelligence zugeordnet ist, verstanden werden kann [L10]. Unter Business Intelligence (BI) werden Techniken und Architekturen bezüglich der Wissensgewinnung, Wissensverwaltung und Wissensverarbeitung zusammengefasst [19].

Für den hier Beschriebenen Anwendungsfall sind die nachfolgend beschriebenen Data Mining Techniken relevant.

2.8.3.1 Entscheidungsbäume

Entscheidungsbäume sind grafische Darstellungsformen von Regelsätzen zur Klassifizierung von Datenobjekten [17]. Ein Entscheidungsbaum besteht aus Knoten und Kanten. Beim obersten Knoten, dem sogenannten Wurzelknoten, startet die Klassifizierung. Die Kanten führen durch bestimmte Regeln bis zu den äußeren Knoten, den sogenannten Blattknoten. Blattknoten stellen die endgültige Klassifizierung dar. Der Wurzelknoten sowie die inneren Knoten stehen jeweils für einzelne Attribute des Datenobjektes, die mit der zugehörigen Aufteilung (dem sogenannten Split) von Kanten eine Regel bilden. Die einzelnen Wege vom Wurzelknoten bis zum Blatt stellen die Regelsätze der Klassifikation dar [17][L21].

Bei der Generierung von Entscheidungsbäumen bzw. deren Regelsätzen wird mithilfe von Kostenfunktionen versucht, den Part mit den niedrigsten Kosten zu finden. Der sogenannte Gini-Index bestimmt den Wert, wie gut ein solcher Part ist [5]. Der Gini-Index wird folgendermaßen berechnet.

$$G = \sum pk * (1 - pk) \tag{1}$$

Die Variable *pk* steht für das Verhältnis zwischen den Gruppen von Eingaben, die durch die Kanten eines Knotens in verschiedene Klassen aufgeteilt werden. Je kleiner *G* ist, desto höher ist die Reinheit einer jeweiligen Klassifizierung durch einen Part des Baumes. *G*=0 steht für die perfekte Reinheit [L21].

Entscheidungsbäume werden im weiteren Verlauf dieser Arbeit für die Unterscheidung von plausiblen und nicht plausiblen Datenattributen angewendet (siehe Abschnitt 3.5.3).

2.8.3.2 Clustering

Beim sogenannten Clustering werden Datenobjekte in Klassen eingeordnet. Im Gegensatz von überwachten Klassifikationsmethoden des Data Minings, wie beispielsweise dem Entscheidungsbaumlernen, gibt es keinen Trainingsdatensatz, bei dem die Daten bereits Klassen zugewiesen sind und diese Zuordnung von dem Verfahren gelernt wird. Beim K-Means Clustering werden metrische (kontinuierliche oder diskrete) Daten anhand einer Abstandsfunktion, im Falle der vorliegenden Arbeit mit der euklidischen Distanz in Klassen eingeteilt. Hierbei wird der Abstand zweier Datenvektoren (welcher die Daten verschiedener Attribute zweier zu vergleichender Datenobjekte enthält) berechnet und die Daten anhand ihrer Abstände hin klassifiziert [75]. Die Funktion der euklidischen Distanz ist in (1) dargestellt.

$$dist_E(a,b) = \sqrt{\sum_i (a_i - b_i)^2} \tag{1}$$

Zunächst ist es beim K-Means Verfahren notwendig, die Anzahl "K" der Klassen im Vorhinein festzulegen, in welche die Daten eingeordnet werden sollen. Im ersten Schritt des Clustering werden die Datenobjekte zufällig in "k" Klassen eingeteilt. Im zweiten Schritt werden für jede dieser Klassen jeweils ein Cluster-Centroid bestimmt. Dies ist ein künstlich erschaffenes Datenobjekt, das aus den jeweiligen Mittelwerten der Werte der Datenobjekte in den jeweiligen Klassen berechnet wird. In Schritt drei werden alle weiteren Datenobjekte nun jeweils in die Klasse desjenigen Cluster-Centroid zugeordnet, welcher dem Datenobjekt am nächsten liegt. Diese Information wird anhand der oben genannten Abstandsfunktion bestimmt. Nun werden die Schritte zwei und drei so lange wiederholt, bis keine Neuzuordnung von Datenobjekten zu Cluster-Centroiden mehr stattfindet [75].

Anschließend sind nun alle Datenobjekte in "k" Klassen eingeteilt. Je höher die Anzahl an Klassen desto ähnlicher sind sich die Werte untereinander. Wäre k gleich der Anzahl der Datenobjekte, so würde jede Klasse nur das Datenobjekt selbst beinhalten.

Angelehnt an Wissuwa et al. (2005) werden in Abbildung 2.7 die verschiedenen Schritte des Clustering grafisch dargestellt [75].

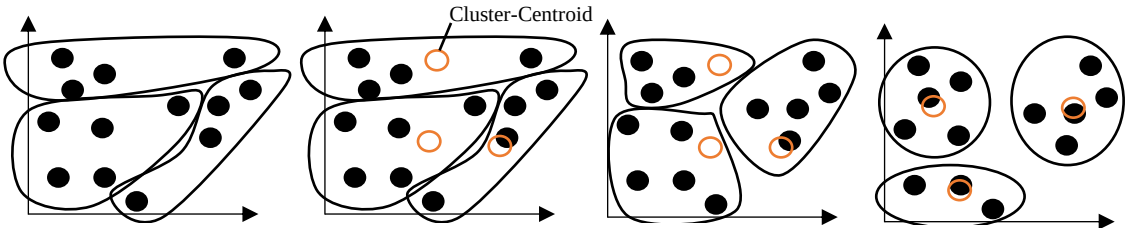


Abbildung 2.7 Durchgeführte Schritte während des Clusterings (in Anlehnung an Wissuwa et al. (2005) [75])

2.8.3.3 Hauptkomponentenanalyse

Das Verfahren der Hauptkomponentenanalyse (Principle Component Analysis, PCA), auch Hauptachsentransformation genannt, wird für die Dimensionsreduktion der Attribute von Datensätzen verwendet. Zusammengefasst ist die PCA eine Methode zur linearen Transformation, damit eine Anzahl von Variablen von einer geringeren Anzahl von Variablen beschrieben werden kann [51][58][37]. Dabei lassen sich Datensätze in einen repräsentativen, niedriger-dimensionalen Raum überführen ohne relevante Informationen zu verlieren [58]. Die durch PCA reduzierten Variablen stellen die Hauptkomponenten dar. Eine jeweilige Hauptkomponente kann dabei durch eine oder mehrere Variablen dargestellt werden [51]. Nach Ng & Soo (2018) lassen sich Hauptkomponenten als die Dimensionen verstehen, entlang derer unsere Datenpunkte am weitesten verteilt sind [51]. Die PCA erzeugt hierbei immer orthogonale (senkrecht zueinanderstehende) Hauptkomponenten. Die sehr umfangreiche Berechnung der Hauptkomponenten der PCA (mittels Bestimmung einer Linearkombination orthogonaler Eigenvektoren der Kovarianzmatrix der Eingabevariablen) wird unter anderem von Dunteman (1989) und von Rehs (2014) beschrieben [25][58]. Diese Art der Dimensionsreduzierung wird in Abschnitt 3.5.2 zur besseren Darstellung von geclusterten Datenelementen durch Dimensionsreduktion angewendet.

2.8.3.4 Text Mining (Kosinus-Distanz)

Text Mining ist ein Teilbereich des Data Mining und beschäftigt sich mit der Analyse von Textdokumenten und Zeichenketten. Da die Inhaltsstofflisten in den Produktdatensätzen der verschiedenen Datenquellen als Zeichenketten vorliegen, dienen Text Mining Verfahren zur Analyse der Ähnlichkeit zwischen Inhaltsstofflisten und Informationsgewinnung. Das nachfolgend beschriebene Verfahren der Kosinus-Distanz-Berechnung findet dabei Verwendung.

Bei der Kosinus-Distanz wird ausgenutzt, dass der Kosinus alle Bedingungen eines Ähnlichkeitsmaßes erfüllt. Bei der unten beschriebenen Kosinus-Distanz-Funktion kann mittels zweier Vektoren ein Winkel errechnet werden, der die Distanz als Winkelgrade zwischen 0° und 90° beschreibt. Hierbei bedeutet 0°, dass zwei Vektoren in allen Bereichen unähnlich zueinander sind und 90°, dass beide Vektoren identisch sind. Der Kosinus-Distanzwert selbst liegt zwischen 0 und 1 und ist somit optimal für die Nutzung als Messwert für die Ähnlichkeit zweier Vektoren. Eine Zeichenkette muss für den Ähnlichkeitsvergleich mittels Kosinus-Distanz-Funktion zunächst als Vektor in einem mehrdimensionalen Vektorraum abgebildet werden [41]. Im Falle der Ähnlichkeit zwischen Inhaltsstofflisten wird hierbei eine Liste I der Inhaltsstoffe (als Zeichenketten) gebildet, die in den zu vergleichenden Produkten vorkommen. Anschließend werden Vektoren gebildet, die die jeweiligen Inhaltsstofflisten repräsentieren. Die Zahl an der Stelle i im jeweiligen Vektor ist 1, wenn der Inhaltsstoff in I an der Stelle i in der Inhaltsstoffliste des Produktes vorkommt. Ansonsten ist es die Zahl 0. Es werden somit zwei gleichlange Vektoren mit Nullen und Einsen gebildet, die anschließend mit der nachfolgenden Funktion verglichen werden können (siehe auch Abschnitt 3.5.1)[L40].

$$dist_C = \cos(\theta) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} X_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \quad (2)$$

2.9 Big Data Anwendungen

Big Data Anwendungen sind in der vorliegenden Arbeit von großer Bedeutung, um bei der großen Zahl von zu bearbeitenden Datensätzen eine ausreichend gute Performance der Bearbeitungs- und Analyseschritte und somit eine hohe Nutzerakzeptanz des zu entwickelnden Systems zu erzielen. Nach Ward & Barker (2013) wird Big Data mit den Begriffen Datenspeicherung und Datenanalyse in Verbindung gebracht [74]. Als Schlüsseldefinition von Big Data wurde demnach die Definition im Bericht von Laney (2001) übernommen [44][74]. Die Definition umfasst die drei Schlüsseleigenschaften Volume (Umfang), Velocity (Geschwindigkeit) und Variety (Vielfalt), die den Begriff Big Data prägen. Insbesondere die ersten beiden Begriffe spielen in der vorliegenden Arbeit eine große Rolle. Wie in späteren Kapiteln beschrieben, werden auch zeitliche Veränderungen der Daten gespeichert, was den Aufwand der Datenanalyse und den Umfang der Datenmenge stark erhöht. Im Analyseprozess werden viele Lebensmitteldatensätze miteinander verglichen und durch Anwendung verschiedener Methoden verarbeitet. Dadurch können in vielen Fällen selbst moderne Rechnerarchitekturen (mit beispielsweise 6-Kern-Prozessor und 64-GB-Hauptspeicher) bei den über 176 000 im FDWH gespeicherten Lebensmitteln mit ihrer Performance schnell an ihre Grenzen kommen und beispielsweise eine zügige Plausibilitätsprüfung neu eingefügter Lebensmitteldaten in Echtzeit

verhindern. Eine solche Prüfung wäre also mit langen Wartezeiten verbunden und würde zu geringer Nutzerakzeptanz von Apps mit Anbindung an das FDWH führen.

2.9.1 TimescaleDB

Um eine wachsende Anzahl von Lebensmitteldaten mit zeitlichen Verläufen performant zu speichern und somit eine hohe Nutzerakzeptanz zu erreichen, wird die Datenbanksoftware TimescaleDB verwendet. TimescaleDB wurde für große Mengen von Zeitreihendatensätzen entwickelt und eignet sich in diesem Feld für die performante Speicherung im Bereich von mehreren Millionen Datensätzen [36][L27]. TimescaleDB vereint die Vorteile von Not-Only-SQL (NOSQL) Datenbanken und den Vorteil von relationalen Datenbanken. Datensätze können relational und somit ohne große Redundanzen übersichtlich strukturiert gespeichert werden. Da das Datenbanksystem, wie die meisten NOSQL-Lösungen, die Daten partitioniert auf einem Verbund eines oder mehreren Rechnern (Rechencluster) abspeichert, wird die zeitliche Performanz des Systems sowie die Skalierbarkeit erhöht [53]. Laut der Webseite der Timescale Inc. erreicht das System eine viel höhere und stabilere Einlese-Rate für Zeitreihendaten, indem es die automatische Zeit-Raum-Partitionierung verwendet und somit das Indizieren und Schreibvorgänge auf einzelne Datenblöcke des Speichers (gegebenenfalls über mehrere Rechnerknoten hinweg) anwendet [L56]. Die Komplexität dieser Prozesse ist für den Anwender nicht sichtbar, diese sehen ausschließlich eine sogenannte Hypertabelle für alle ihre Daten. Eine Hypertabelle ist speziell für sogenannte Chunks (in viele einzelne Tabellen partitionierte Daten) strukturiert, verhält sich aber dennoch so wie eine normale SQL-Tabelle [L56]. Timescale Inc. gibt an, dass mit TimescaleDB Anwender Millionen von Datenpunkten pro Sekunde auf einen einzigen Rechner schreiben können, bei einer Verteilung über mehrere verteilte Rechnerknoten sogar mehrere zehn Millionen Datenpunkte pro Sekunde. Des Weiteren sind Speichereinsparungen von 90-95% möglich [L56].

2.9.2 Das Apache Spark Framework in der Programmiersprache Python (PySpark)

Laut der offiziellen Webseite von Apache Spark [L50] handelt es sich bei dem Framework um eine einheitliche Engine für die Analyse großer Datenmengen bzw. für die Ausführung von Data Engineering, Data Science und maschinellem Lernen auf Einzelknotenrechnern oder mehreren Rechnerknoten. Das Framework ist in verschiedenen Programmiersprachen geschrieben, in der vorliegenden Arbeit wird es als Python-Framework (PySpark genannt) angewendet. PySpark verteilt die Daten im Hauptspeicher mehrerer Clusterpartitionen eines Rechenclusters, wendet darauf

Operationen an und führt die Ergebnisse zusammen [80][E]. Im Detail werden die Daten als sogenanntes Resilient Distributed Dataset (RDD, übersetzt: elastischer verteilter Datensatz) [L49] gespeichert. Programmierte Datenanalyseoperationen werden auf dem RDD ausgeführt, dabei wird das RDD von Entwicklern als einzelne Datentabelle betrachtet, obwohl dieses auf mehrere Hauptspeicher in verschiedenen Rechenclustern aufgeteilt ist. Die Speicherung, Verteilung der Daten, die Anwendung der programmierten Operationen sowie die Zusammenführung der Ergebnisse werden vom Apache Spark Framework verwaltet [50]. Durch das verteilte Verarbeiten der Daten im Hauptspeicher des Rechenclusters wird die Echtzeitverarbeitung der Daten enorm beschleunigt [80][E].

2.10 Methoden der Validierung

Zur Beurteilung von Datenanalysemethoden kann eine Wahrheitsmatrix dazu dienen die Wahrscheinlichkeiten zur Genauigkeit und Präzision sowie die Richtig-Positiv-Rate (RPR) und die Falsch-Positiv-Rate (FPR) zu ermitteln [40]. Diese Wahrscheinlichkeiten dienen in der vorliegenden Arbeit zur Bewertung der angewendeten Ähnlichkeits- und Plausibilitätsanalyse (siehe Abschnitt 3.8.2). Tabelle 2.4 zeigt den Aufbau der hier genutzten Wahrheitsmatrix.

Plausibilitätsmessung:		**Tatsächliche Plausibilität**	
		plausibel	implausibel
Plausibilitätstest	Positiv (Test positiv)	Richtig-Positiv (RP)	Falsch-Positiv (FP)
	Negativ (Test negativ)	Falsch-Negativ (FN)	Richtig-Negativ (RN)

Tabelle 2.4 Wahrheitsmatrix zur Berechnung von Genauigkeit, Präzision, RPR und FPR

Die vier Felder der Wahrheitsmatrix enthalten zum einen den Wert Richtig-Positiv (RP) mit der Anzahl aller durch die Plausibilitätsanalyse richtig als plausibel eingestuften Testdaten. Der Wert Falsch-Positiv (FP) beschreibt die Anzahl aller falsch als plausible eingestuften Testdaten. Des Weiteren beschreiben

die Werte Falsch-Negativ (FN) und Richtig-Negativ (RN) die fälschlicherweise als implausibel bzw. richtig als implausibel eingestuften Testdaten. Zusätzlich wird die Richtig-Positiv-Rate (RPR) berechnet, die das Verhältnis von richtig als plausibel eingestuften Testdaten zu den als gesamt als plausibel eingestuften Testdaten bestimmt. Die Falsch-Positiv-Rate (FPR) beschreibt das Verhältnis von falsch als plausibel eingestuften Testdaten zu den gesamt als plausibel eingestufte Testdaten [29][13][2]. Die Funktionen zur Berechnung von RPR und FPR werden in (1) und (2) aufgeführt.

$$RPR = \frac{RP}{RP+FN} \tag{1}$$

$$FPR = \frac{FP}{RN+FP} \tag{2}$$

Zwei weitere statistische Kenngrößen dienen der Einschätzung der Präzision PREC (3) und der Genauigkeit ACC (4) des Plausibilitätsverfahrens.

$$PREC = \frac{RP}{RP+FP} \tag{3}$$

$$ACC = \frac{RP+RN}{RP+FP+RN+FN} \tag{4}$$

PREC bestimmt das Verhältnis der Menge an richtig als plausibel eingestuften Testdaten zu der Menge der insgesamt als plausibel eingestuften Testdaten. ACC bestimmt das Verhältnis richtig (als plausibel und implausibel) eingestufter Testdaten im Verhältnis zu den Testdaten insgesamt [29][13][2].

Neben der Bestimmung von Präzision und Genauigkeit können durch Bildung der sogenannten Receiver Operating Characteristic (ROC) anhand der Werte RPR und FPR Kenntnisse über die Güte der Klassifizierung gewonnen werden. Die Koordinaten (RPR, FPR) werden in ein Koordinatensystem mit dem Intervall [0;1] für die X- und die Y-Achse eingetragen. Wie gut durch das Analysemodell klassifiziert wurde, kann man nun an den eingetragenen Punkten ablesen. Der ROC-Graph ist eine Diagonale die vom Punkt (0,0) bis zum Punkt (1, 1) reicht (siehe Abbildung 2.9). Die linke Seite dieser Diagonalen zeigt die gültigen Klassifizierungen, die rechte Seite die ungültigen, bei denen keine sinnvolle Klassifizierung möglich ist. Entlang der Diagonale liegen gleich viele richtige und falsche Klassifizierungen. Je mehr richtige Klassifizierungen, also je näher der Punkt (RPR, FPR), desto besser wird durch das Netz klassifiziert. Der Punkt (0, 1) steht für die perfekte Klassifizierung, wie Abbildung 2.8 zeigt [29].

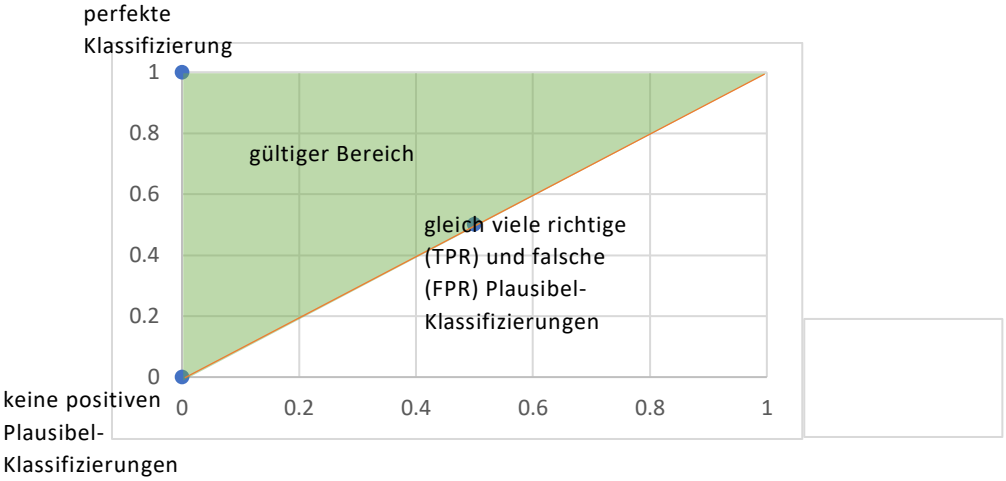


Abbildung 2.8 ROC Diagramm mit Diagonalen und Beschreibung der Bereiche

3 Ergebnisse

3.1 Angewendete Entwicklungsumgebungen und Werkzeuge

Je nach Anwendungszweck wurden für die Entwicklung des FDWH, des ETL-Prozesses und der weiteren Analyse- und Datenverarbeitungsprozesse verschiedene Lösungen zur Implementierung und Software-Werkzeuge angewendet.

Als Datenbanksystem wurde die auf große Zeitreihen spezialisierte Datenbanksoftware TimescaleDB verwendet (siehe Abschnitt 2.9.1). Der gesamte ETL-Prozess sowie die Ähnlichkeits- und Plausibilitätsanalyse und einzelne Data Profiling Module (die während des ETL-Prozesses ausgeführt werden) wurden in der Programmiersprache Python entwickelt. Durch Zuhilfenahme des Programmierframeworks Pandas, welches ebenso in Python geschrieben ist, können performante Algorithmen auf sogenannten Dataframes angewendet werden. Ein weiteres in Python geschriebenes Framework für die Anwendung sehr großer Datenmengen (Big Data) ist PySpark (siehe Abschnitt 2.9.2) [50][80].

Die Suchfunktionen des FDWH, die API und die webbasierte Benutzerschnittstelle wurden mithilfe des Spring Boot Frameworks in der Programmiersprache Java entwickelt. Spring Boot ist leicht konfigurierbar und beinhaltet fortgeschrittene Programmieransätze des Spring Frameworks. Beispielsweise erlaubt der Ansatz der Dependency Injection das konfigurierbare Einbinden und Austauschen von Softwaremodulen. Durch aspektorientierte Programmierung (AOP) ist das Programmieren von Aspekten (z. B. Sicherheit oder Logging) unabhängig anderer Klassen im Framework erlaubt. Diese Ansätze vereinfachen die Modularisierung und Zusammensetzung von Anwendungen aus verschiedenen einfachen Java-Klassen [67][76][77]. Durch Anwendung des Software-Pattern Model-View-Control (MVC) können in Spring Boot sowohl Datenmodell und Persistenz, Benutzerschnittstelle und grafische Visualisierung als auch die Anwendungslogik durch Klassen voneinander getrennt werden [68][78]. Als Server-Umgebung für die Spring Boot Anwendungen (API und Webtechnologien) dient ein für Java-Server-Technologien entwickelter Tomcat Server [72].

Für einen Teil des Data Profilings (beispielsweise bei der Überprüfung von Datentypen in externen Datenquellen), dem Clustering und dessen Visualisierung sowie der Evaluierung wurde die Data Mining Software KNIME angewendet. Die Zeichenkettenanalyse wurde mit der Data Profiling Software TOS-DQ durchgeführt.

3.2 Anwendung von Data Profiling Methoden

Zur Transformation der Daten externer Datenquellen in ein einheitliches Format werden Data Profiling Methoden verwendet, mit denen die Eigenschaften von Daten genauer untersucht werden. Des Weiteren können durch Data Profiling Regeln aufgestellt werden, durch die die verschiedenen Elemente eines Datensatzes auf Qualitätsmängel überprüft werden können. Typische Qualitätsmängel sind hierbei die nachfolgend aufgeführten [10]:

- Fehlerhafte (inkorrekte) Daten
- Inkonsistente (zueinander widersprüchliche) Daten
- Duplikate (doppelte Daten)
- Uneinheitlich repräsentierte Daten
- Veraltete Daten
- Irrelevante Daten (für den jeweiligen Anwendungszweck)
- Unverständliche Daten (bedingt durch qualitativ mangelhafte Metadaten)

Als Beispiel dient Abbildung 3.1, welche die Qualitätsmängel zweier Auszüge unterschiedlicher Datenquellen mit Lebensmittelproduktdaten aufzeigt.

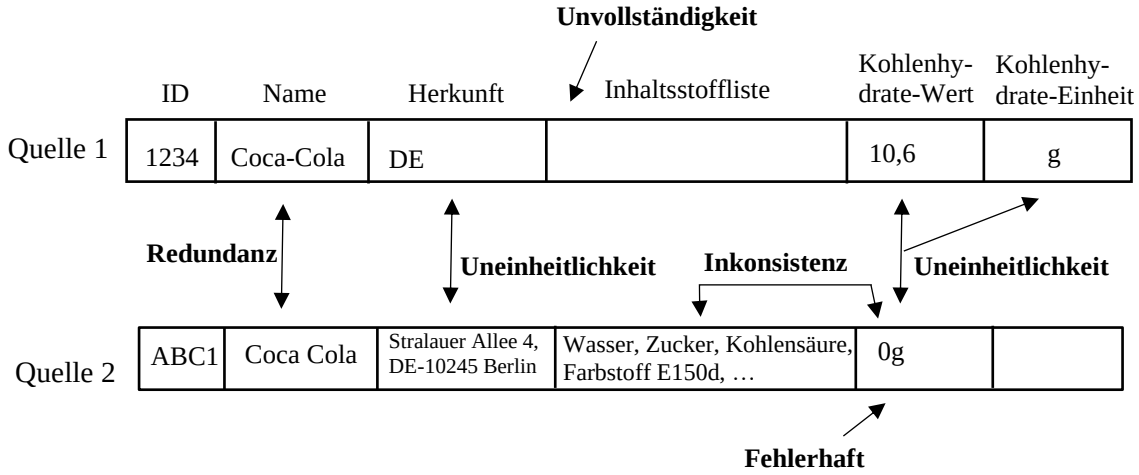


Abbildung 3.1 Beispiel von Qualitätsmängeln zweier Auszüge unterschiedlicher Datenquellen mit Lebensmittelproduktdaten

Abbildung 3.1 zeigt das ein Lebensmittelprodukt mit dem Namen Coca-Cola doppelt vorhanden ist und sich in der Schreibweise des Produktnamens unterscheidet. Die Herkunft ist uneinheitlich, einmal mit vollständiger Adresse und einmal nur mit dem jeweiligen Ländercode angegeben. Die Inhaltsstoffliste fehlt ist in Quelle 1 und ist somit fehlerhaft. Die Information bezüglich des Nährwertes für Kohlenhydrate wird zudem uneinheitlich angegeben. Bei Quelle 2 steht die Einheit der Angabe direkt im Feld für den Wert und bei Quelle 1 steht diese in einem separaten Feld. Der Wert 0g ist bei dem angegebenen Produkt fehlerhaft (richtig wäre 10,6g). Außerdem steht der Wert 0 im Widerspruch mit der Inhaltsstoffliste. Da Zucker im Produkt enthalten ist, müssen zwingend auch Kohlenhydrate enthalten sein. Deshalb muss die Angabe als inkonsistent eingestuft werden.

Der nachfolgende Abschnitt erläutert zunächst die Analyse spezifischer Beziehungen unter den verschiedenen Datenattributen, die sogenannte Ontologie der Datenattribute. Anschließend werden Methoden der Metadatenanalysen und die Verwendung externer Wissensdaten vorgestellt. Ziel ist die Informationsgewinnung von Attributeigenschaften, die zur Ermittlung fehlender Werte oder zur Vereinheitlichung von Werten dienen. Durch die anschließend dargestellten Datentyp- und Musteranalysen konnten Regeln erstellt werden, die beschreiben, in welcher Form die Datensätze im FDWH gespeichert werden.

3.2.1 Ontologie der Datenattribute

Ontologien sind spezielle Beschreibungen von Beziehungen zwischen Datenattributen, die meist in der Form von Verbindungen ähnlich der Satzbau-Semantik zwischen Subjekten und Objekten dargestellt sind, wobei der Name der Verbindung dem Prädikat entspricht. Solche Beschreibungen können durch Graphen gebildet werden, indem jeweils Subjekte und Objekte als Knoten und Prädikate, die als Beschreibungen der Beziehungen fungieren, als Kanten zwischen den Knoten dargestellt werden. Solche Ontologien werden beispielsweise im semantischen Web zur Beschreibung von reellen Datenobjekten verwendet [6][69].

Die Abbildung 3.2 enthält, entsprechend der Darstellungsmethode in Dengel et al. (2012), den Graphen eines Ontologie-Modells, welches durch eine systematische Analyse von Metadaten der externen Datenquellen sowie weiterer Informationen über Lebensmitteldaten (unter anderem Informationen auf Lebensmittelproduktverpackungen und in der Literatur über Lebensmittel, beispielsweise in Biesalski et al. (2020) [12]), entstanden ist (vgl. Abschnitt 2.8.1) [23]. Hierbei wurden die für das FDWH relevanten Datenobjekte identifiziert. Ferner wurden Zusammenhänge und Beziehungen zwischen den Datenobjekten sowie dessen Eigenschaften und Datentypen ermittelt.

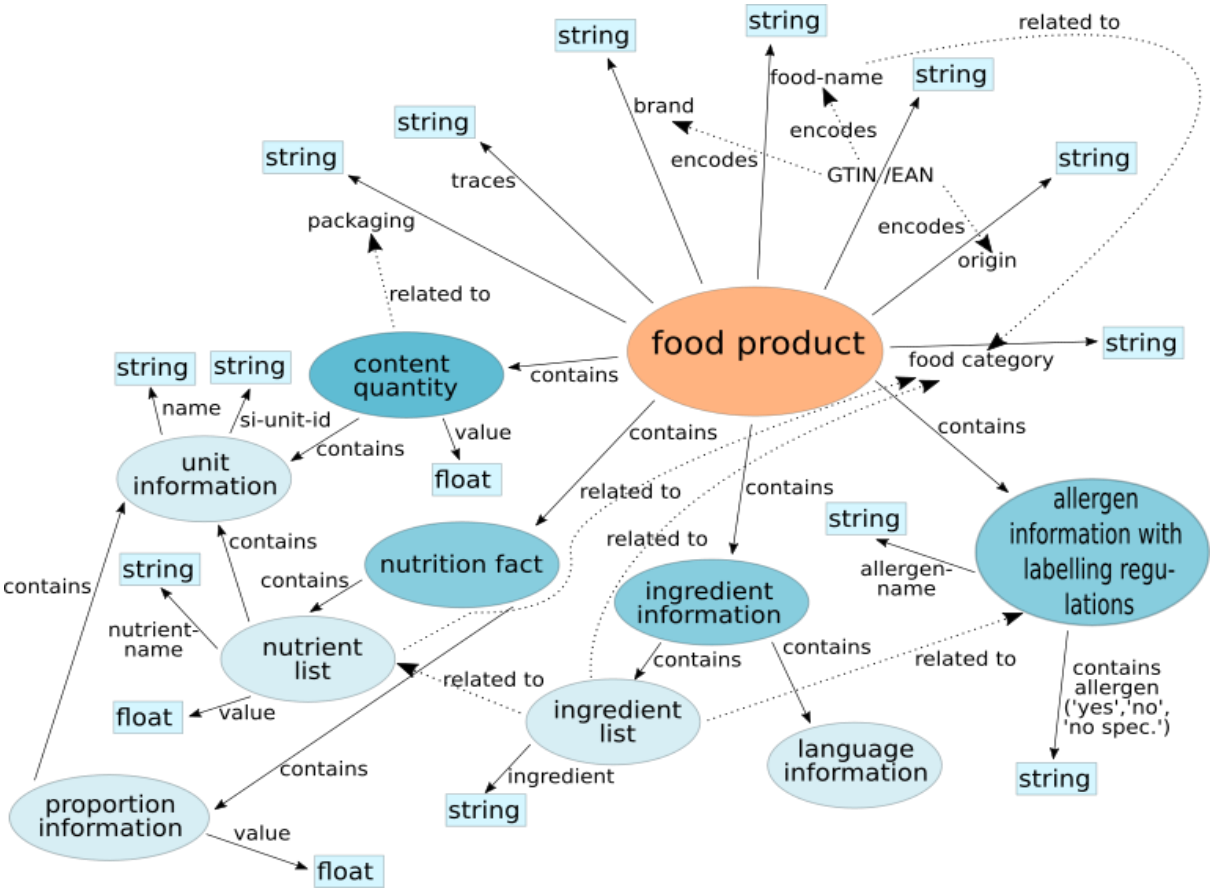


Abbildung 3.2 Ontologie-Modell der Datenelemente des FDWH

Die Knoten im Graphen stellen jeweils Datenobjekte bzw. Datentypen dar. Die Kanten stehen zusammen mit der jeweiligen Kantenbezeichnung für die Art von Beziehungen (und dessen Bedeutung) zwischen den Knoten. Darüber hinaus werden durch die Kanten Objekteigenschaften dargestellt, die beispielsweise eine Relation eines Datenobjektes zu einem bestimmten Datentyp beschreiben. Die Beziehung *contains* beschreibt beispielsweise, welches Datenobjekt einem anderen Objekt zugeordnet ist bzw. welche Eigenschaften eines Datenobjektes in einem anderen Datenobjekt enthalten sind. Zum Beispiel enthält das Datenobjekt, welches die Inhaltsstoffliste (*ingredient list*) beschreibt, die einzelnen Informationen über einen jeweiligen Inhaltsstoff (*ingredient information*). Die Beziehung *related to* besagt, dass wenn sich eine Information eines Datenobjektes ändern würde, dies Auswirkung auf die Informationen eines weiteren Datenobjektes hätte. Würde sich beispielsweise die Inhaltsstoffliste eines Lebensmittelproduktes (bezeichnet als *food product*) ändern, z. B. die Zutat Zucker würde entfernt, so würde sich auch die Liste der Nährwerte (*nutrient list*) ändern, da sich Anzahl an Kohlenhydrate im Lebensmittelprodukt reduzieren würde. Des Weiteren könnte eine Änderung der Inhaltsstoffliste auch eine Änderung der Informationen über Allergene (*allergen information with labelling regulations*) erwirken.

Das beschriebene Ontologie-Modell wurde für die Erstellung der Datenbank-Architektur des FDWH, dem Aufbau des ETL-Prozesses sowie der Analyse der Datenqualität durch Data Profiling und Ähnlichkeitsanalyse-Methoden genutzt, da die speziellen Verbindungen zwischen Datenobjekten leicht ersichtlich werden. Außerdem enthalten die Metadaten des FDWH Informationen, die durch das Ontologie-Modell dargestellt sind. Diese Informationen sind behilflich bei der Transformation von Daten externer Datenquellen während des ETL-Prozesses.

3.2.2 Metadatenanalyse

Durch eine Analyse von Metadaten, der externen Datenquellen und der in Abschnitt 3.1.1 dargestellten Ontologie, fand eine Ermittlung notwendiger Eigenschaften über Struktur und Datentypen dieser bereitgestellten Daten statt. Diese Eigenschaften liefern nützliche Informationen für die nachfolgend beschriebenen Datentyp- und Musteranalysen. Die Metadaten wurden entweder von den Betreibern der verschiedenen Datenlieferanten bereitgestellt oder wurden aus Datenbankschemata herausgelesen. Weitere Metadaten wurden für die Entwicklung von Transformationsregeln während des ETL-Prozesses eingesetzt (siehe Abschnitt 3.3.3). Beispielsweise konnten Metadaten der OpenFoodFact.org-Datenbank Informationen dazu liefern, dass dort alle Nährwerte für Kohlenhydrate, Fett und Proteine in der Einheit Gramm gespeichert werden. Da das FDWH die erwähnten Nährwerte in der Einheit Milligramm speichert, wird anhand der Metadateninformation die Umrechnung der Daten von Gramm zu Milligramm getriggert, was den Importprozess deutlich vereinfacht. Zusätzlich können während des Importierens von Daten ins FDWH Berechnungen von fehlenden Nähr- bzw. Energiewerten durchgeführt werden. Beispielsweise enthalten Metadaten Informationen zum Energiegehalt von Nährwerten (z. B. entsprechen jeweils 1 Gramm Kohlenhydrate und 1 Gramm Proteine 4,1 Kilokalorien und 1 Gramm Fett entspricht 9,3 Kilokalorien Energie) die dafür herangezogen werden können.

Neben den bereits genannten Metainformationen wird weiteres Wissen für die Datenverarbeitungsprozesse des FDWH zur Verfügung gestellt. Dies sind beispielsweise Listen mit Bezeichnern und dazugehörigen Synonymen von Lebensmittelinhaltsstoffen, die Anhand von Internetrecherchen und mithilfe der Synonym-Datenbank openthesaurus.de erstellt wurden [L31]. Solche Listen dienen zur Vereinheitlichung von Inhaltsstoffbezeichnern vor der Ähnlichkeitsanalyse (beschrieben in Abschnitt 3.5.1) [E]. Des Weiteren dient zu einer solchen Vereinheitlichung eine Liste mit allen E-Nummern, die in Inhaltsstofflisten vorkommen und deren Bezeichnern. Eine weitere eigens erstellte Liste mit Kategorie-Bezeichnern, die (bei Bedeutungsgleichheit) einen Verweis jeweils auf die Kategorien des BLS und die Kategorien der WikiFood-Datenbank enthält, dient zur späteren automatisierten Kategorisierung der Lebensmitteldatensätze (siehe Abschnitt 3.4) [B].

3.2.4 Datentypanalyse

Da sich die Datenstruktur der Lebensmitteldaten in ihren verschiedenen heterogenen Datenquellen unterscheidet, wird zur Vereinheitlichung dieser Datensätze die Datentypanalyse angewendet, damit keine zusätzlichen Inkonsistenzen entstehen [3]. Beim Laden eines Wertes in die Zieltabelle des FDWH muss dieser Wert dem Typ der Tabellenspalte entsprechen, in die der Wert eingefügt werden soll. Hierzu wird der Typ des Wertes zunächst ermittelt und es wird versucht, falls dieser nicht dem Zieltyp entspricht, den Wert in den richtigen Typ umzuwandeln. Zur Ermittlung von Datentypen (falls diese nicht aus Metadaten der jeweiligen Quelldatenbank ausgelesen werden können, z. B. bei CSV-Dateien ohne zugehörige Schemabeschreibung), dient die freie Open-Source-Software Konstanz Information Miner (KNIME) [L29]. KNIME erkennt den Datentyp einer Tabellenspalte, sofern alle Daten in der Spalte dem gleichen Typ zugeordnet werden können. Sind beispielsweise alle Werte einer Spalte Fließkommazahlen bis auf einen Wert, der ein Wort als Zeichenkette darstellt, so nimmt KNIME den Typ String (für die Zeichenkettenspeicherung) an, da Fließkommazahlen ebenso als Zeichenkette gespeichert werden können. Die Umwandlung eines Datentyps wird sowohl durch Boardmittel der jeweiligen Programmiersprache versucht (sogenannten Cast-Operatoren) wie auch mittels eigener programmierter Algorithmen. So kommt es beispielsweise vor, dass in den Datenquellen Fließkommazahlen mit dem Komma (“,“) als Trennzeichen versehen sind, obwohl diese für den Cast-Operator der Programmiersprache einen Punkt (“.“) zur Typumwandlung enthalten müssten. Gegebenenfalls muss vor der Verwendung des Cast-Operators eine Umwandlung von Komma in Punkt erfolgen. Anschließend kann der Wert durch den Cast-Operator richtig zugeordnet werden [L62]. Welchem Datentyp die verschiedenen Attribut-Felder der FDWH-Datenbank zugeordnet sind, wird in Tabelle 3.1 aufgelistet. Des Weiteren enthält die Tabelle eine Beschreibung des jeweiligen Attributfeldes.

Attributfeld	**Beschreibung**	**Datentyp**
food_name	Lebensmittel-/Produktname	varchar/String (Zeichenkette)
food_name_search	Zeichenkette des Produktnamens für einfachere Selektierung in der Datenbank (Produktname in Großbuchstaben, Sonderzeichen entfernt, Leerzeichen als Unterstrich “_“ codiert	varchar/String
food_name_soundex	Sound-Expression-Code des Produktnamens	varchar/String
brand_name	Markenname	varchar/String

brand_name_search	Zeichenkette für bessere Selektierung (siehe food_name_search)	varchar/String
brand_name_soundex	Sound-Expression-Code des Markennamens	varchar/String
origin_name	Herkunftsland	varchar/String
gtin_number	GTIN	varchar/String
ingredient_list	Inhaltsstoffliste	varchar/String
language_id	Codierung der Sprache der Inhaltsstoffliste nach ISO-639-1	varchar/String
language_name	Name der Sprache der Inhaltsstoffliste	varchar/String
energy_kcal	Nährwert für Energie in Kilokalorien (kcal)	double (Gleitkomma-zahl)
carbohydrates	Nährwert für Kohlenhydrate (in Milligramm)	double
fat	Nährwert für Fett (in Milligramm)	double
proteins	Nährwert für Proteine (in Milligramm)	double
content_value	Werte der Inhaltsmenge	double
si_unit_id	Codierung der SI-Einheit der Inhaltsmenge	varchar/String
si_unit_name	Name der SI-Einheit	varchar/String
gl	Wert der aussagt, ob Gluten enthalten ist (j=ja, n=nein, k=keine Angabe)	char/charakter (Zeichen)
la	Wert der aussagt, ob Laktose enthalten ist	char/charakter
additional_info	Zusätzliche Informationen	varchar/String
category	Kategoriebezeichnung	varchar/String

Tabelle 3.1 Attributfelder der FDWH-Datenbank inklusive Datentyp Zuordnung

In Tabelle 3.1 sind nur vier der wichtigsten Makronährstoffe-Attribute aufgelistet. Die Datenbanktabelle enthält insgesamt 154 verschiedene Attribute mit Mikro- bzw. Makronährstoffen, Vitaminen und Mineralstoffen, die aufgrund einer besseren Übersicht nicht aufgeführt sind. Diese enthalten alle denselben Datentyp für Gleitkommazahlen (*double)*. Des Weiteren sind der Übersicht halber auch nur die Attribute der Allergene Gluten (gl) und Lactose (la) aufgeführt anstelle aller 14 häufig auftretenden Allergene. Diese enthalten alle den Datentyp Charakter (*char*).

3.2.5 Musteranalyse in Zeichenketten

Durch die Musteranalyse wird die Struktur von Zeichenketten in den Daten der verschiedenen Datenquellen untersucht. Hierbei wird untersucht, ob die Zeichenkette eines Attributes immer dieselbe Struktur oder unterschiedliche Strukturen besitzt und welche Art von Zeichen (ob große oder kleine Buchstaben, Sonderzeichen, numerische Zeichen, etc.) enthalten sind. Außerdem wird untersucht, ob die Länge der Zeichen eines Attributes immer gleich oder unterschiedlich ist und wie kurz die kürzeste bzw. wie lang die längste Zeichenkette ist. Nach einer Untersuchung, ob diese Ermittelten Strukturen der externen Datenquellen für eine konsistente und vereinheitlichte Darstellung der jeweiligen Daten im FDWH taugen, wurden Regeln für die Speicherung solcher Daten im FDWH festgelegt, nach denen Zeichenketten jeweils vor der Speicherung transformiert werden müssen.

Die Data Profiling Software mit dem Namen „Talend Open Studio for Data Quality“ (kurz: TOS-DQ, der Firma Talend) erlaubt eine Darstellung solcher Muster. Die Software wandelt dann beispielsweise zur besseren Darstellung alle Buchstaben in das Zeichen a um und alle Zahlen in die Zahl 9. Somit wird beispielsweise aus der Zeichenkette „Pizza Salami“ (die einen Produktnamen darstellt) das Muster „aaaaa aaaaaa“. Anschließend kann man durch TOS-DQ anzeigen lassen, welche Muster in den Daten wie oft vorkommen und welche keine Muster, also keinen Wert (Empty field) enthalten (siehe Abbildung 3.3). Beispielsweise kommt das Muster „Aaaaaaaaaa“ (Produktname besteht aus zehn Buchstaben) in den Daten des Attributes der Datenquelle OpenFoodFacts.org in 644 Fällen vor. In 2828 Fällen ist kein Wert enthalten.

Die erhaltenen Muster aller Zeichenketten der Daten eines bestimmten Attributes wurden unter einem jeweiligen regulären Ausdruck [L16][L51] zusammengefasst und für die Attribute und Datenquellen in Tabelle 3.2 dargestellt.

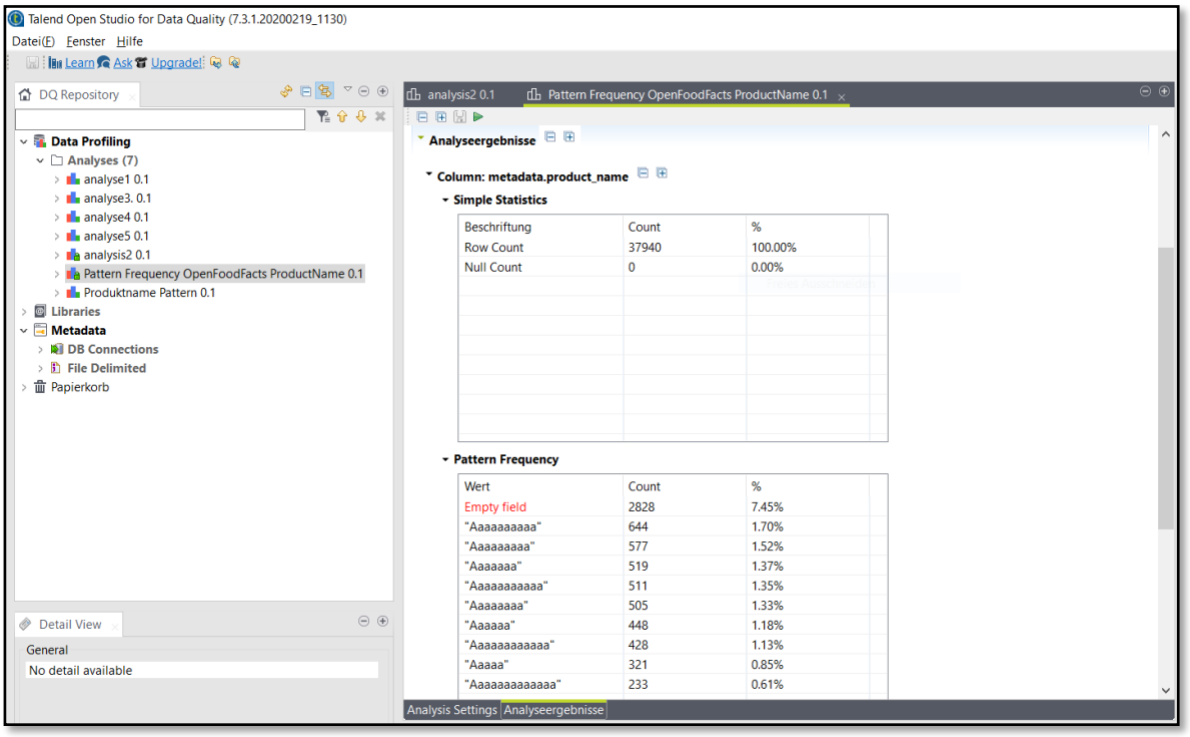


Abbildung 3.3 Beispiel einer Zeichenkette-Musteranalyse in TOS-DQ

	BLS	**Wiki-Food**	**Open-Food-Facts**	**Food-Repo**	**das-ist-drin**	**Danone**
Produktname	.+	.+	.+| Ø	.+	.+	.+
Markenname	/	.+| Ø	.+| Ø	.+| Ø	.+	/
Herkunft	/	.+| Ø	.+| Ø	.+| Ø	Ø	/
EAN/GTIN	/	[0-9]{8,13}| Ø	[0-9]{8,13}| Ø	[0-9]{8,13}| Ø	[0-9]{8,13}	Ø
Kennzeichnungspflichtige Allergene	/	.+| Ø	.+| Ø	Ø	Ø	/
Sonstige Informationen	/	.+| Ø	.+| Ø	.+| Ø	Ø	Enthält Gelatine | Frei von Gelatine
Inhaltsstoffliste	/	.+| Ø	.+| Ø	.+| Ø	.+| Ø	.+

Mengenangabe	/	/	/	/	/	/
Nährwertinformationen	/	/	/	/	/	/

Tabelle 3.2 Ermittelte reguläre Ausdrücke der Speicherung von Zeichenketten in den externen Datenquellen

Wie man in der Tabelle sehen kann, unterliegen beispielsweise die Produktnamen aller Datenquellen, außer der Quelle OpenFoodFacts.org, dem regulären Ausdruck „.+“. Dies bedeutet, dass ein beliebiges Zeichen (durch „.“ dargestellt) ein- oder mehrmals (durch „+“ dargestellt) vorkommt. Bei OpenFoodFacts.org kommt es vor, dass die Zeichenkette für Produktname keinen Wert enthält. Dies wird im regulären Ausdruck durch das Zeichen „Ø“ dargestellt. Der gesamte Ausdruck für den Produktnamen in OpenFoodFacts.org lautet: “.+|Ø“ (ein beliebiges Zeichen, das ein- bzw. mehrmals vorkommt oder der Wert ist nicht besetzt). Wenn unter einem Attribut eines bestimmten Datensatzes der Wert nicht vorhanden oder nicht als Zeichenkette enthalten ist, so wird der Tabelleneintrag mit dem „/“ gekennzeichnet.

Aus den regulären Ausdrücken aus Tabelle 3.2 wurde untersucht, welche Struktur die Zeichenketten im zentralen FDWH haben müssen, um die oben erwähnte einheitliche Darstellung der Werte aus allen Datenquellen zu erlauben und gleichzeitig inkonsistente oder fehlerhafte Werte zu vermeiden. Dieses Ergebnis wird in Tabelle 3.3 dargestellt. Tabelle 3.4 enthält eine Beschreibung aller verwendeten regulären Ausdrücke aus Tabelle 3.3. Der Produktname eines Produktes kann z. B. in jeder der Datenquellen ein oder mehrere beliebige Zeichen beinhalten. Diese Regel wird so übernommen. Das Attribut Produktname sollte im FDWH immer einen Wert beinhalten und niemals den leeren Wert, weil dies sonst eine genaue Identifikation des Produktes erschweren oder unmöglich machen würde. Ein weiteres Beispiel ist der Datenwert für das Attribut GTIN/EAN. Eine korrekte GTIN bzw. eine korrekte EAN beinhaltet immer zwischen 8 und 13 numerische Zeichen. Deshalb ist der reguläre Ausdruck für dieses Attribut im FDWH entsprechend festgelegt. Allergeninformationen liegen, wenn überhaupt, in den Datenquellen als Freitext vor. Beispielsweise gibt die WikiFood.eu-Datenquelle an, in welchen Produkten bestimmte kennzeichnungspflichtige Allergene nicht enthalten sind. Z. B. wird dann ein Produkt, welches keine Nüsse enthaltet, mit „nussfrei“ deklariert. Sonstige Informationen sind, wie bereits erläutert, Informationen über Verpackung des Produktes oder zusätzliche Informationen über den Produktinhalt (z. B., ob es sich um ein veganes Produkt handelt).

	FDWH
Produktname	.+
Markenname	.+| Ø
Herkunft	/
EAN/GTIN	[0-9]{8,13}| Ø
Kennzeichnungspflichtige Allergene	/
Zusätzliche Produktinformationen	.+| Ø
Inhaltsstoffliste	\w+.+[, ;:.-_]+\w+.+| Ø
Mengenangabe	/
Nährwertinformationen	/

Tabelle 3.3 Ermittelte reguläre Ausdrücke für die einheitliche Speicherung von Zeichenketten im FDWH

Regulärer Ausdruck	**Beschreibung**
.+	Ein oder mehrere beliebige Zeichen
Ø	Leerer Wert
.+| Ø	Ein oder mehrere beliebige Zeichen oder leerer Wert
\w+.+[, ;:.-_]+\w+.+| Ø	Ein oder mehrere beliebige alphanumerische Zeichen („\w+“), dann ein oder mehrere beliebige Zeichen, danach ein oder mehrere Trennzeichen (Komma, Leerzeichen, Semikolon, Doppelpunkt, Punkt, Bindestrich oder Unterstrich), dann wieder ein oder mehrere alphanumerische und ein oder mehrere beliebige Zeichen (da mindestens zwei Inhaltsstoffe vorhanden sein müssen, damit es als Inhaltsstoffliste zählt und nicht nur das Produkt selbst in der Liste enthalten ist, muss mindestens ein Trennzeichen enthalten sein).
[0-9]{8,13}| Ø	Die Ziffern 0 bis 9 zwischen 8- und 13-mal hintereinander oder leerer Wert
/	Keine Werte in Datenquelle vorhanden oder Datentyp nicht als Zeichenkette enthalten bzw. Wert entstammt aus standardisierter Tabelle (z. B. bei Angabe der Herkunft)

Enthält Gelatine | Frei von Gelatine	Entweder der Wert „Enthält Gelatine“ oder der Wert „Frei von Gelatine“

Tabelle 3.4 Beschreibung der regulären Ausdrücke aus Tabelle 3.3

Die regulären Ausdrücke in Tabelle 3.3 dienen somit als Regeln für die Qualitätsanalyse während des Transformationsprozesses. Wenn Daten einem solchen Ausdruck nicht entsprechen, werden diese wenn möglich umgewandelt oder aussortiert.

3.3 Zusammenführung der Lebensmittel-Datenquellen in ein zentrales Data-Warehouse

3.3.1 Datenbankschema

Die relationale Datenbank des FDWH unterliegt dem Snowflake-Schema. Dieses beschreibt die Struktur relationaler Datenbanktabellen, die eine zentrale Tabelle enthält, von der aus allen weiteren Tabellen mit ihren Relationen nach außen hin abzweigen. Die zentrale Tabelle steht somit in einer 1:n Relation (ein Datensatz der zentralen Tabelle steht in Relation mit mehreren Datensätzen der abgezweigten Tabelle) mit weiteren Tabellen, die wiederrum mit weiteren Tabellen in einer 1:n Relation stehen. Diese Verbindungen können sich beliebig fortsetzten. Die Struktur ähnelt der einer Schneeflocke, daher der Name des Schemas. Jede Ebene des Snowflake-Schemas, bei der sich Datentabellen in weitere Relationen abzweigen, wird auch Dimension genannt. Die zentrale Datenbanktabelle bildet hierbei die erste Dimension. Eine grafische Darstellung des Snowflake-Schemas erfolgt durch Abbildung 3.4. Die Verwendung des Snowflake-Schemas bringt einige Vorteile mit sich. Durch die übersichtliche Struktur sind die einzelnen Tabellen leichter zu pflegen. Des Weiteren ermöglicht das Schema performantere Abfragen im Vergleich zu vollständig normalisierten Datenmodellen. Durch die Aufteilung der verschiedenen Dimensionen und die Verbindung durch 1:n Beziehungen können die Teilinformationen eines Datensatzes auf der jeweiligen Dimension ohne das Bilden großer verschachtelter und verknüpfter Abfragen (z. B. mit sogenannten „joins“ in der SQL-Abfragesprache) durchgeführt werden, was viele der Abfragen im FDWH performanter macht [45][30]. Einzelne Informationen der Lebensmitteldaten im FDWH können somit schneller abgefragt werden. Ein Nachteil ist, dass das Schema nur in Teilen normalisiert ist und somit mehr Redundanzen als vollständig normalisierte Datenbanktabellen enthält [11]. Durch die Kombination des Schemas mit der in Abschnitt 2.9.1 beschriebenen Anwendung der TimescaleDB, werden Techniken von NOSQL Datenbanken, zur schnellen und skalierbaren

Datenverarbeitung, in Kombination mit den Vorteilen der besseren Strukturmöglichkeiten durch SQL und dem Snowflake-Schema angewendet.

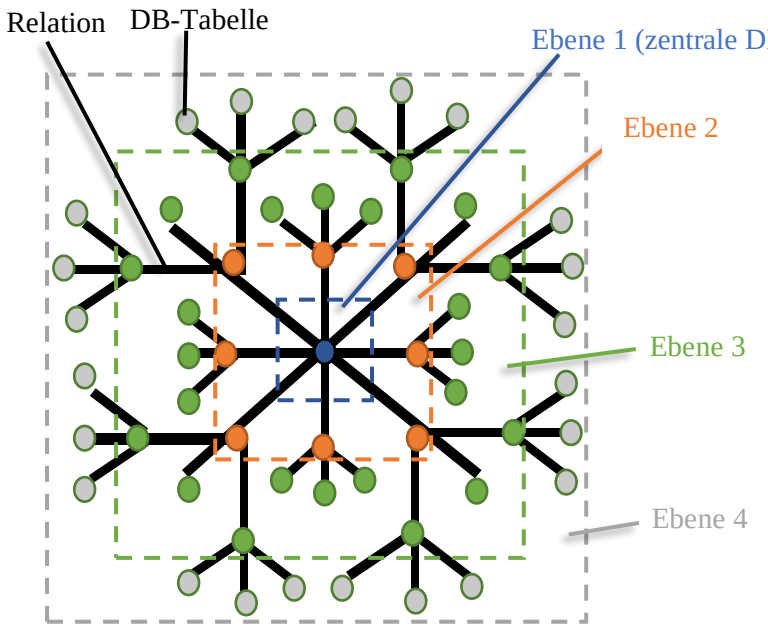


Abbildung 3.4 Grafische Darstellung des Snowflake-Schema

Die Datenbanktabellen orientieren sich größtenteils an dem in Abschnitt 3.2.1 beschriebenen Ontologie-Modell und den Datenbankschemata der externen Datenquellen, sofern diese übermittelt wurden. Die zentrale Tabelle food_condition bildet die zeitliche Dimension des FDWH. Diese fungiert als sogenannte Faktentabelle [30] mit Verknüpfungen zu den Tabellen mit Lebensmittelinformationen und beinhaltet als Zeitstempel das Datum und die Zeit, wann der Datensatz in das FDWH eingefügt wurde. Wenn sich die Informationen eines Lebensmittels ändern, wird ein neuer Datensatz mit aktuellem Zeitstempel angelegt. Auf diesem Weg wird die Änderungshistorie der Lebensmitteldaten im FDWH dokumentiert. Somit sind Änderungen an der Zusammensetzung von Lebensmitteln sowie der Vergleich von Datensätzen für Datenqualitätsprüfungen nach einer Veränderung von Attributwerten auch in Zukunft noch möglich [52]. Die zweite Dimension des FDWH-Schemas unterteilt sich in eine Dimension der Lebensmittelinformationen, eine Dimension der Lebensmittelkategorien (Verknüpfung von Lebensmitteldatensätzen zu den jeweiligen Kategorien) und eine Dimension mit der Faktentabelle latest_food mit den jeweils aktuellen Versionen der einzelnen Lebensmitteldatensätzen. Diese Tabelle enthält somit anstelle der Änderungshistorie von Lebensmitteldatensätzen immer nur eine Kopie des aktuellen Datensatzes eines jeden Lebensmittels. Die dritte FDWH-Dimension enthält Tabellen mit weiteren Informationen zu den Lebensmitteln (zum Teil Wissensdaten wie die gespeicherte Liste mit SI-Einheiten oder Sprachcodes und Informationen zu der ursprünglichen Datenquelle) und weiteren Informationen zu der Kategorisierung (Haupt- und Unterkategorien, Maßeinheiten zur verzehrten

Lebensmittelmenge, verknüpft mit den Kategorien). Abbildung 3.5 enthält eine grafische Darstellung des gesamten FDWH-Schemas mit den farblich markierten Dimensionen.

Abbildung 3.5 Darstellung des gesamten FDWH-Schema

3.3.2 Generierung der Datenbanktabellen-IDs

Die Datenbanktabellen mit Informationen zu den Lebensmitteln in der zweiten FDWH-Dimension enthalten als Primärschlüssel jeweils eine Identifikationsnummer (ID), die jeweils einen Hashwert beinhaltet, der aus den Lebensmittelinformationen der jeweiligen Tabelle generiert wird. Tabelle 3.5 beschreibt, welche Informationen zur Generierung einer bestimmten ID herangezogen wurden. Beispielsweise wurde der Hashwert der ID der Tabelle content, welche Informationen über die Inhaltsmenge eines Lebensmittels enthält, aus dem Wert des Attributes content_value (Wert der Inhaltsmenge) und des Attributes si_unit_id (Kennung der SI-Einheit der Inhaltsmenge) gebildet. Beide Werte wurden zu einer Zeichenkette zusammengesetzt bevor der Hashwert generiert wurde.

ID der Datenbank	**Attribute für Hashwertgenerierung**
brand	brand_name_search
origin	origin_name
gtin	gtin_number
ingredient_list	ingredient_list
nutrition_fact	Alle Tabellenattribute mit Ausnahme der ID selbst
content	content_value, si_unit_id
allergen_info	Alle Tabellenattribute mit Ausnahme der ID selbst
additional_info	additional_info
data_source_info	Alle Tabellenattribute mit Ausnahme der ID selbst

Tabelle 3.5 Informationen zur Generierung der FDWH-IDs

Als Funktion für die Hashwertberechnungen wurde das Secure Hash Verfahren SHA-256 Verfahren ausgewählt, welches Hashwerte der Länge von 256 Bit berechnet und mit einer großen Sicherheit gewährleistet, dass keine zwei Datensätze mit unterschiedlichen Informationen den gleichen Hashwert erhalten. Gleichzeitig wird eine für die Speicherung in einer relationalen Datenbank vernünftige Länge der Hashwert-Zeichenkette beibehalten [L7].

Die ID (food_id) der Faktentabelle food_condition wird durch den Hashwert aus den Kombinationen des Lebensmittelnamens und der IDs der Tabellen brand, origin und content gebildet. Diese Werte wurden gewählt, da Lebensmittelprodukte durch die Kombination aus Lebensmittelname, Markenname, Herkunft und Inhaltsmenge eindeutig identifiziert werden, auch wenn die GTIN des Produktdatensatzes

nicht gesetzt ist. Die ID von food_condition in FDWH-Dimension 1 ist gleichzeitig die ID der Tabelle food (die Lebensmittelname und Informationen zur Produktsuche enthält) und der Tabelle latest_food.

Die Primärschlüssel der Zeitdimensionstabelle food_condition werden durch den Zeitstempel insertion_time und die generierte food_id erstellt. Ändert sich die Zusammensetzung eines Lebensmittels, bleibt der Wert von food_id gleich, der Zeitstempel ändert sich aber. Je nach Art einer solchen Änderung ändert sich gleichzeitig eine (oder mehrere) der IDs in den Tabellen ingredient_list, nutrition_facts und allergen_info (siehe Abbildung 3.6).

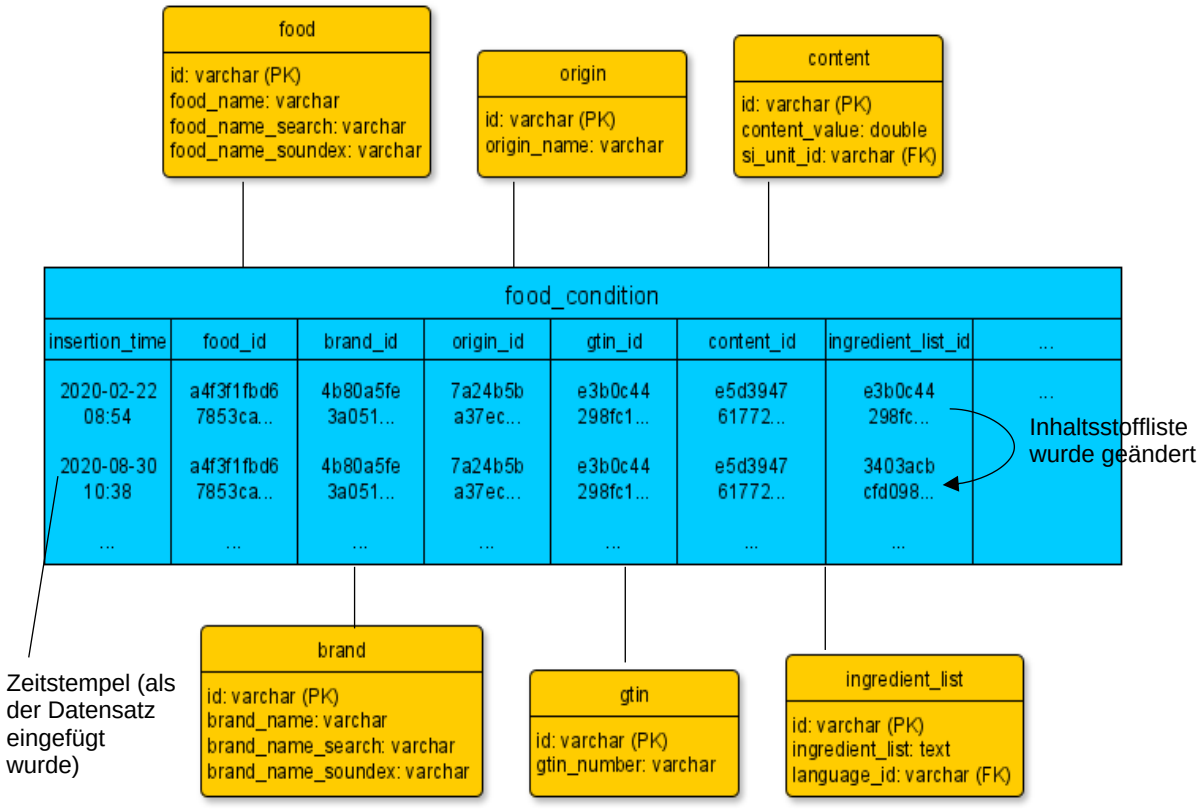


Abbildung 3.6 Auszug aus der FDWH-Datenbank als Beispiel für die ID-Änderung bei gleichzeitiger Änderung in der Zusammensetzung des Lebensmittelproduktes

3.3.3 Entwicklung des ETL-Prozesses

Wie bereits beschrieben, liegen Datensätze der externen Lebensmitteldatenquellen in verschiedenen Formaten wie beispielsweise JSON-Dateien, Excel-Dateien bzw. Datenbank-Dumps von MongoDB oder PostgreSQL Datenbank vor. Der Extrahierungs-Part des ETL-Prozesses wurde so entwickelt, das Lebensmitteldatensätze aller Quellen zunächst in sogenannte Dataframe geladen werden. Ein Dataframe ist in diesem Fall eine virtuelle Datentabelle, welche mit Hilfe des Pandas-Programmierframeworks der Programmiersprache Python gebildet wird [L54]. In Pandas entspricht ein Dataframe einer virtuellen Datentabelle auf die Operationen zur Bearbeitung der Daten durchgeführt werden können. Abbildung 3.7 zeigt als Beispiel die Struktur eines Dataframes, welches mit der Funktion *print* im Python-Quellcode ausgegeben wurde. Pandas stellt vordefinierte Funktionen wie die in (1) gezeigte Funktion *transform* zur Verfügung, durch die Daten im Dataframe performant verarbeitet werden. In (1) ist ein Beispiel dargestellt, in dem mithilfe der *transform*-Funktion die eigens implementierte Funktion *profileGtins* auf alle Werte des Attributes *gtin_number* (enthält die GTIN des Produktdatensatzes) des Dataframes *gtinNumberDf* angewendet wird. Die Funktion *profileGtins* überprüft, ob die jeweilige GTIN dem entsprechenden regulären Ausdruck entspricht.

gtinNumberDf['gtin_number'] = gtinNumberDf['gtin_number'].transform(profileGtins) (1)

	id	name
0	762	Frucht & Honig Sanddorn
1	763	Fruchtdessert Rote Grütze
2	764	Frucht Pur 75% Aprikose
3	765	Frucht Pur 75% Erdbeere
4	766	Frucht Pur 75% Heidelbeere
5	767	Frucht Pur 75% Himbeere
6	768	Frucht Pur Rote Johannisbeere
7	769	Amaranth-Früchte-Müsli
8	770	Amaranth Früchte Müsli
9	771	Lavendelblütenhonig

Abbildung 3.7 Beispiel der Struktur eines Dataframes im Pandas-Framework

Weiterhin werden bei dem Ladevorgang und dem anschließenden Transformationsvorgang Metadaten mit Informationen zu den Datenattributen der externen Datenquellen aus sogenannten INI-Dateien ausgelesen. INI-Dateien eignen sich gut zur Speicherung von Konfigurationsdaten. Solche Daten können in Sektionen und Key-Value-Paare aufgeteilt und gespeichert werden. Python stellt geeignete Frameworks zur Verfügung, die eine Verarbeitung der Konfigurationsdateien erleichtern. Ändern sich

Attribute in den Datenquellen oder kommen neue Datenquellen hinzu, so können die INI-Dateien leicht angepasst bzw. neu angelegt werden. Dies erlaubt zum Teil eine dynamische Entwicklung des ETL-Prozesses. Beispielsweise sind in den INI Dateien die Attributnamen für Produkt- und Markenname der jeweiligen Datenquellen gespeichert. Durch diese Information weiß der ETL-Prozess, in welchen Datenfeldern der Datenquelle er die gewünschten Informationen findet. Die Datensätze werden in den jeweiligen Dataframes transformiert und somit in eine einheitliche und konsistente Struktur bzw. ein standardisiertes Format umgewandelt [A][C]. Einige der Datenfelder beispielsweise mit den im Abschnitt 2.5.4 genannten Allergeninformationen erschweren eine dynamische Anpassung des ETL-Prozesses. Dies liegt an den sehr unterschiedlichen Strukturen der zugrunde liegenden Informationen. Ob eines der 14 Auslöser im Produkt enthalten (oder nicht enthalten) ist, wird durch eine unterschiedliche Benennung der Auslöser bzw. der Art der Darstellung der Information ausgedrückt. In solchen Fällen müssen spezielle Funktionen auf den Dataframes implementiert werden, die die jeweiligen gespeicherten Informationen in ein angepasstes Format formatieren. Das Anwenden dieser Funktionen auf das jeweilige Dataframe geschieht mithilfe der oben genannten *transform*-Funktion. Im Falle der Allergeninformationen gibt es demnach spezifisch für die jeweiligen Datenquellen entsprechende Funktionen, die diese Daten transformieren. Bei Hinzufügen einer neuen Datenquelle kann eine solche Funktion implementiert und durch *transform* in den Programmcode des ETL-Prozesses inkludiert werden. Neben der Anwendung der oben genannten INI-Dateien werden Wissensdaten mit Informationen über ISO-639-1 Sprachcodes, ISO-3166-1 Ländercodes, Geodaten, Synonyme von Inhaltsstoffen, kennzeichnungspflichtige Allergene, Kategorisierungen und E-Nummern-Listen als Wissensdaten für die Transformationen bereitgestellt. Die Dataframes werden nach der Transformation in die jeweiligen Tabellen des FDWH geladen. Abbildung 3.8 stellt den gesamten ETL-Prozess grafisch dar.

```
[OFF]
file_path = \02-Promotion\10-Datequellen\ETL Data\
file_name = 20201125_off_important_data.csv
default_basic_attributes = food_name;data_source_food_id;food_insertion_date;food_modification_date;brand_name;origin_name;gtin_number;
original_basic_attributes = product_name;_id;created_t;last_modified_t;brands;countries_tags;code;
default_content_attributes = content_value;
original_content_attributes = product_quantity;
default_info_attributes = additional_info;
original_info_attributes = packaging;
default_category_attributes = category_from_datasource;
original_category_attributes = categories;
default_allergen_attributes = lu;sw;se;sf;sl;nu;en;la;so;wt;fi;ei;kr;gl;
```

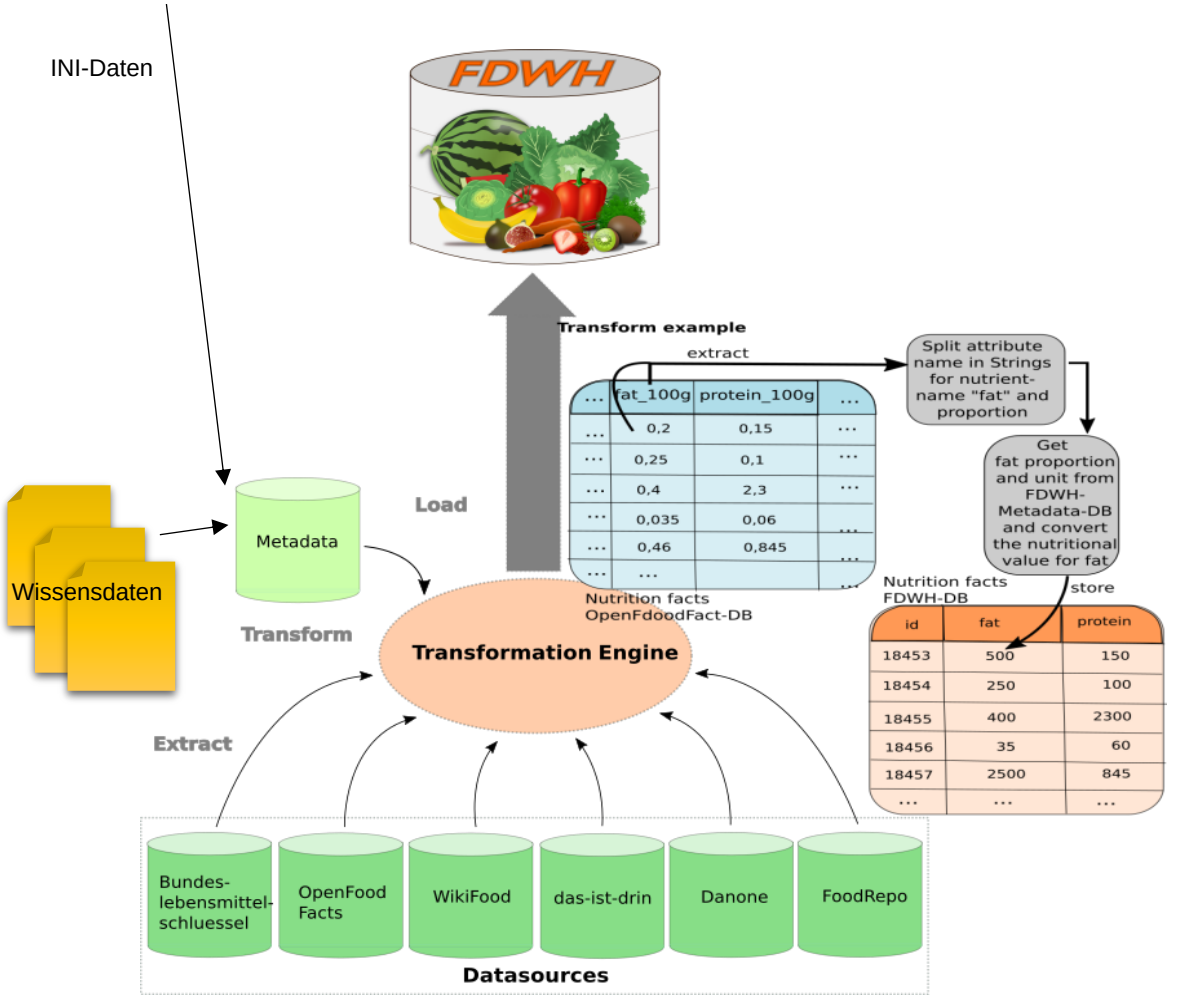


Abbildung 3.8 Grafische Darstellung des ETL-Prozesses

3.4 Automatisierte Kategorisierung der Lebensmittelproduktdaten im FDWH

Mithilfe von Ernährungsberaterinnen des deutschen Allergie- und Asthmabundes (DAAB) wurde eine Kategorisierung von Lebensmitteln entwickelt. Diese Kategorisierung beschreibt die Produktart (z. B. Reisgerichte, Produkte mit Rindfleisch, etc.) und dient der geordneten Darstellung und erweiterten Suche von Lebensmittelproduktdaten. Des Weiteren wurde durch die Kategorisierung eine Zusammenfassung von Lebensmitteldaten für Auswertealgorithmen in elektronischen Ernährungstagebüchern im DiDiER-Projekt ermöglicht. Zusätzlich wurden die Kategorien mit gleichbedeuteten Kategorien der Bundeslebensmittelschlüssel- und der WikiFood-Datenbank gemappt. Anhand des Mappings können viele der Daten von den beiden genannten Anbietern automatisiert mit Kategorien versehen werden [B]. Während der Transformation des ETL-Prozesses wird zusätzlich der Lebensmittelproduktname mit Kategorien verglichen. Bei Übereinstimmung von Teilen des Produktnamens wird jeweils die Kategorie für das Produkt übernommen. Außerdem können bei fehlender Kategorisierung die Kategorien von ähnlichen Produkten übernommen werden, die durch die Ähnlichkeitsanalyse ermittelt wurden.

Die oben genannte Art der automatisierten Kategorisierung eines Produktes (nach Beschreibung der Produktart) ist besonders bei Anwendungszwecken sinnvoll, bei denen die Zusammensetzung der Inhaltsstoffe eines Produktes von Bedeutung ist. Dies wäre beispielsweise in Bezug zu Nahrungsmittelunverträglichkeiten oder -allergien der Fall, zum Beispiel in elektrischen Ernährungstagebüchern, wenn gegebenenfalls durch eine Kategorie Produkte gruppiert werden sollen deren Produkte oder die Inhaltsstoffe von Produkten einem Allergen entsprechen. Eine weitere Art der Kategorisierung ist die des Clustering-Verfahrens (beschrieben in Abschnitt 2.8.1 und angewendet in 3.5.2). Sie führt zu sinnvollen Aussagen über die Nährwertzusammensetzung des Produktes, da klar ist, welcher Cluster welchen Nährwertbereichen entspricht. Innerhalb des FDWH sind beide beschriebenen Arten der Kategorisierung etabliert.

3.5 Ähnlichkeits- und Plausibilitätsanalyse

Wie bereits erwähnt, reichen die dargestellten Data Profiling Methoden nicht aus, um fehlerhafte oder fehlende Daten zu erkennen bzw. zu ermitteln. Mit der in diesem Kapitel vorgestellten Ähnlichkeitsanalyse sollen ähnliche oder gleiche Datensätze durch Vergleichen von Attributwerten festgestellt werden. Anschließend wird eine Gesamtähnlichkeit und ein weiterer Wert ermittelt, der zur Bewertung der Plausibilität eines Datensatzes verwendet wird.

Es besteht die Annahme, dass, wenn zwei Lebensmittel ähnlich oder gleich zueinander sind, diese Ähnlichkeiten bzw. Gleichheit in der Produktbezeichnung sowie in der Lebensmittelzusammensetzung aufweisen. Es wird des Weiteren vermutet, das fehlende Werte eines Lebensmitteldatensatzes in einem weiteren Datensatz übernommen werden können, wenn eine Ähnlichkeit der beiden Datensätze besteht. Als Beispiel wird das Lebensmittelprodukt Erdbeerjoghurt zweier unterschiedlicher Marken X und Y betrachtet. Die Daten über die Zusammensetzung der beiden Produkte sind nachfolgend aufgelistet.

- Bezeichnung (Name) Produkt 1: X Erdbeerjoghurt
  - Inhaltsstoffliste: Joghurt mild 3,5 % Fett, Zucker, Erdbeeren, Erdbeermark, modifizierte Stärke, färbendes Rote Bete-Saftkonzentrat, natürliches Aroma, Säureregulatoren: (Natriumcitrate, Apfelsäure).
  - Nährwerte:
    - Fett 3 000 mg
    - Kohlenhydrate 13 000 mg
    - Proteine 2 900 mg

- Bezeichnung (Name) Produkt 2: Y Joghurt mit Erdbeere:
  - Inhaltsstoffliste: Joghurt mild, Erdbeeren, Zucker, Glukose-Fruktose-Sirup, modifizierte Stärke, färbendes Rote Bete-Saftkonzentrat, Aroma.
  - Nährwerte:
    - Fett 2 800 mg
    - Kohlenhydrate 14 500 mg
    - Proteine 3 000 mg

Wie bei den beiden Datensätzen zu erkennen ist, weisen diese Ähnlichkeiten im Produktnamen, bei der Inhaltsstoffliste und den Nährwerten auf, deren Werte nahe beieinander liegen. Fehlt einer der Werte in einem Datensatz, so wäre die Übernahme des Wertes aus dem jeweils anderen Datensatz möglich. Würden beispielsweise die Nährwerte des Datensatzes von Produkt 2 fehlen, so könnten anhand der Kenntnis, dass die Produkte ähnlich sind, die Nährwerte von Produkt 1 übernommen werden. Eine

solche Kenntnis wird dann durch die Ähnlichkeitsanalyse der beiden vorhandenen Werte der Attribute für Produktbezeichnung und Inhaltsstoffliste gewonnen. Umgekehrt können auch Inhaltsstofflisten durch die Ähnlichkeitsanalyse von Produktbezeichnung und Nährwertebereichen vervollständigt werden. Der Wert bezüglich der Bezeichnung des Produktes (der Produktname) ist nach den Regeln des ETL-Prozesses nie unbelegt.

Durch initiale Annahme und spätere Bestätigung bei der Auswertung der Ähnlichkeitsanalyse wurden die bedeutenden Erkenntnisse erzielt, dass zur Klassifikation der Ähnlichkeit eines Lebensmittels zwei der drei Attributbereiche Produktbezeichnung (Produktname), Inhaltsstoffliste und die Nährwerte ausreichend sind. Im Bereich der Nährwerte sind die Werte für Fett, Kohlenhydrate und Eiweiß zusammen mit der Produktbezeichnung für die Ähnlichkeitsanalyse ausreichend, da diese Werte zusammen das Produkt ausreichend gut charakterisieren. Die Angaben der Nährwerte wurden im Zuge des ETL-Prozesses bezüglich der genaueren Darstellung in Milligramm (mg) umgerechnet. Allgemein muss bei der Ähnlichkeitsanalyse zwischen der Analyse von Zeichenketten und der Analyse von numerischen Werten unterschieden werden. Produktbezeichnungen und Inhaltsstofflisten liegen als Zeichenkette vor. Nährwerte werden durch numerische Werte dargestellt. In den nachfolgenden Unterkapiteln werden die jeweiligen Verfahren der Ähnlichkeitsanalyse je nach Art der Analyse erläutert [F]. Anschließend erfolgt die Analyse der Plausibilität der Datenwerte in den verglichenen Lebensmitteln sowie die Beschreibung der Bestimmung einer Gesamtähnlichkeit des Lebensmittels aus den ermittelten Ähnlichkeitswerten. Wichtig ist hierbei die Untersuchung, inwiefern die automatisierten Verfahren zur Qualitätskontrolle und -optimierung beitragen und inwiefern sich diese Art und Weise in der Performanz gegenüber der manuellen bzw. händischen Qualitätskontrolle auswirkt.

3.5.1 Ähnlichkeitsanalyse von Zeichenketten

Im ersten Schritt der Analyse müssen die Zeichenketten für die Analyse der Ähnlichkeit in eine standardisierte Form gebracht werden. Im ersten Schritt werden alle alphabetischen Großbuchstaben der jeweiligen Zeichenkette in Kleinbuchstaben umgewandelt. Des Weiteren werden alle Sonderzeichen und numerischen Zeichen entfernt. In einem weiteren Schritt werden Stoppwörter aus den Zeichenketten gelöscht. Stoppwörter sind sogenannte Füllwörter (wie beispielsweise Artikel, Konjunktionen etc.), die keine Relevanz über die Aussage des Inhaltes der Zeichenkette besitzen und in diesem Falle keine nützliche Information zum Datensatz liefern. Des Weiteren werden Verben und Adjektive anhand einer Liste entfernt. Während des ETL-Prozesses wurden, wie bereits beschrieben, Synonyme von Inhaltsstoffen zur Vereinheitlichung und E-Nummern durch deren Bezeichnungen ersetzt. Nach dieser Transformation werden Mengen mit den jeweiligen Wörtern der zu vergleichenden Zeichenketten gebildet.

Im obigen Beispiel sähen diese Mengen der Inhaltsstofflisten von Produkt 1 (Menge A) und Produkt 2 (Menge B) folgendermaßen aus.

A= {joghurt, fett, zucker, erdbeeren, erdbeermark, stärke, betesaftkonzentrat, aroma, säureregulatoren, natriumcitrate, apfelsäure} (1)

B= {joghurt, erdbeeren, zucker, glucose, fructose, sirup, stärke, betesaftkonzentrat, aroma} (2)

Im nächsten Schritt wird eine Vereinigungsmenge (Menge V) aus den Mengen A und B gebildet.

V= {joghurt, fett, zucker, erdbeeren, erdbeermark, stärke, betesaftkonzentrat, aroma, säureregulatoren, natriumcitrat, apfelsäure, glucose, fructose, sirup} (3)

Die beiden Mengen A und B werden nun in Vektoren anhand der folgenden Regeln umgewandelt.

$$\forall\, v_i \in V, \quad i = 1,2,\dots,|V|, \quad x_i = \begin{cases} 0, & v_i \in A \\ 1, & v_i \notin A \end{cases} \tag{4}$$

$$\forall\, v_i \in V, \quad i = 1,2,\dots,|V|, \quad y_i = \begin{cases} 0, & v_i \in B \\ 1, & v_i \notin B \end{cases} \tag{5}$$

Im Beispiel entstehen die beiden nachfolgend aufgeführten Vektoren.

x = [1,1,1,1,1,1,1,1,1,1,1,0,0,0] (6)

y = [1,0,1,1,0,1,1,1,0,0,0,1,1,1] (7)

Mit der anschließenden Formel der Kosinus-Distanz-Berechnung wird die Ähnlichkeit der beiden zu vergleichenden Zeichenketten gebildet [L40].

$$z = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \approx 0{,}6 \tag{8}$$

Der Ähnlichkeitswert der Produkte 1 und 2 liegt somit bei ca. 0,6 und besagt eine Ähnlichkeit der Zeichenketten von 60 %.

Nachfolgend in Abbildung 3.9 ein weiteres Beispiel der Ähnlichkeitsanalyse zweier Produktbezeichnungen „Coca-Cola Light ohne Zucker“ und „Coca-Cola Zero 0 % Zucker“.

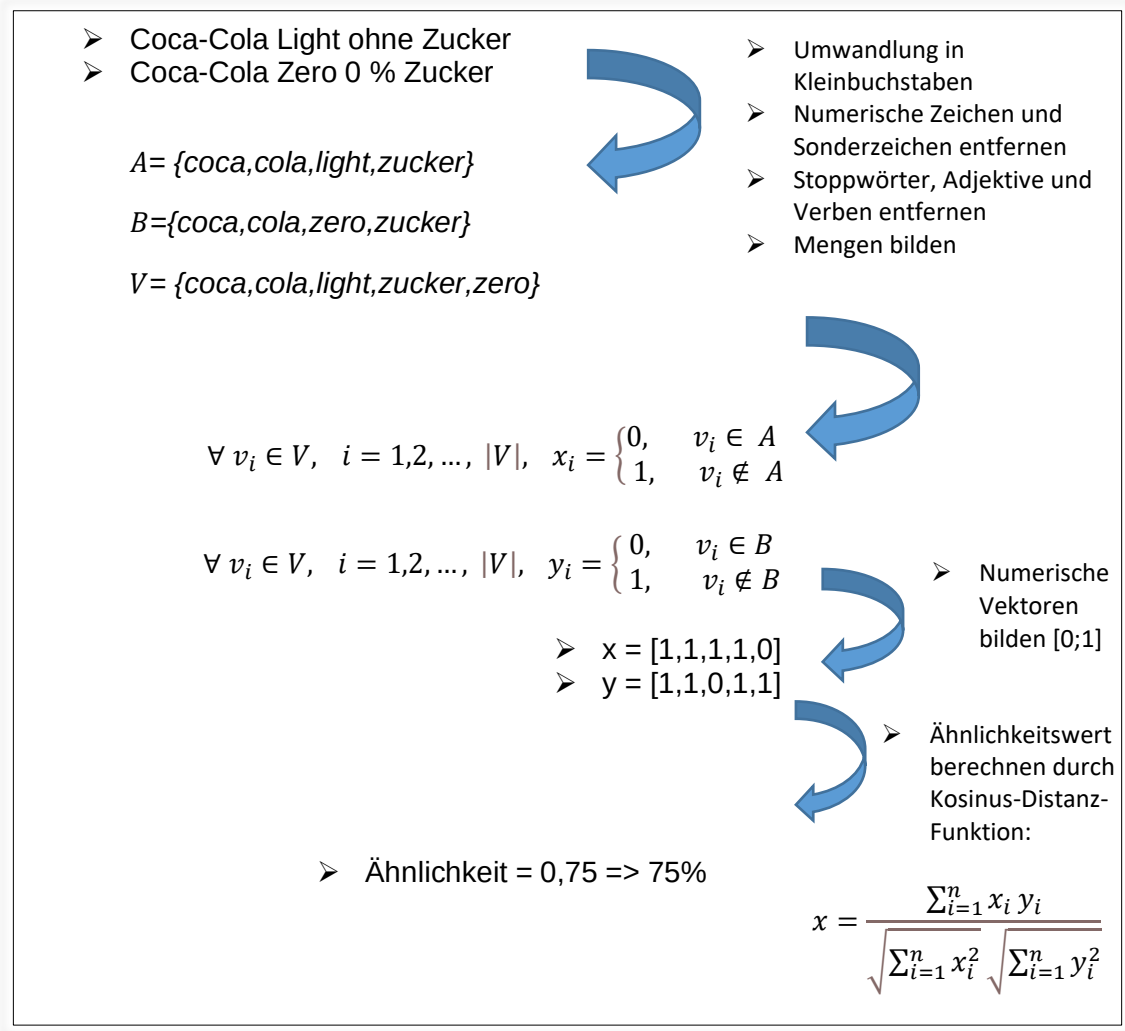


Abbildung 3.9 Beispiel der Ähnlichkeitsanalyse zweier Produktbezeichnungen

3.5.2 Ähnlichkeitsanalyse von numerischen Werten

Eine Ähnlichkeit besteht im Bereich der Nährwerte Fett, Kohlenhydrate und Proteinen (jeweils in mg angegeben), wenn die Nährwerte zweier Produkte nah beieinander liegen. Beispielsweise liegen im obigen Beispiel (Abschnitt 3.4.1) der beiden Produktdatensätze Produkt 1 und Produkt 2 die Nährwerte für Fett jeweils im Bereich zwischen 2800 und 3000 mg.

Zur Ähnlichkeitsanalyse der numerischen Nährwerte und gleichzeitig zur Bestimmung, inwieweit diese Werte auseinander liegen dürfen, wurde das in Abschnitt 2.8.3 beschriebene Clustering-Verfahren K-Means angewendet. Dieser Algorithmus ist ein guter Ansatz, um Attribute zu klassifizieren, deren Werte nahe beieinander liegen. Hierbei werden anhand der Nährwerte von Produktdatensätzen und mithilfe der euklidischen Distanzfunktion Klassen gebildet, in welche die Produktdatensätze anhand ihrer

Nährwertebereiche eingeordnet werden. Während der Ähnlichkeitsanalyse können diese Klassen (sogenannte Cluster) zweier Datensätze miteinander verglichen werden, um eine Ähnlichkeit zu bestimmen. Die Schwierigkeit bei der Verwendung von K-Means Clustering liegt in der Festlegung der Anzahl der Klassen (Anzahl K), in die die Datensätze unterteilt werden sollen. Mit Hilfe von K wird gleichzeitig die Größe der Wertebereiche bestimmt, innerhalb derer sich die Nährwerte der Produktdatensätze innerhalb einer Klasse befinden. Ist der Wertebereich zu groß, es wurden somit zu wenige Klassen gebildet, besteht keine ausreichend große Ähnlichkeit der Nährwerte. Eine anschließende Bewertung der Gesamtähnlichkeit und Plausibilität eines Produktdatensatzes ist nicht mehr möglich. Bei zu vielen Klassen und zu kleinen Wertebereichen werden oftmals Ähnlichkeiten zwischen Lebensmitteln nicht mehr erkannt. Auch dies erschwert die Bestimmung der Gesamtähnlichkeit sowie der Plausibilität eines Produktdatensatzes. Demnach musste vor der Anwendung des K-Means Clustering dem Ziel nachgegangen werden, wie K im vorliegenden Anwendungsfall bestimmt werden kann. Dazu lieferte zunächst der von Fink (2019) beschriebene Ansatz eine Orientierung [L15]. Hier wurden Datensätze des Anbieters OpenFoodFacts.org in 20 Klassen anhand der Nährwerte Fett, Kohlenhydrate und Proteine einsortiert. Anschließend wurden die gebildeten Klassen den Labels zugeordnet, die den Produktdatensätzen von Anwendern der Plattform OpenFoodFacts.org zugewiesen wurden. Die Labels entsprechen eigenen Kategorien der Anwender. Hiermit sollte unter anderem festgestellt werden, ob diese zugewiesenen Labels einer jeweiligen Klasse untereinander ebenfalls ähnlich sind. Dies war beim Großteil der Labels der Fall, beispielsweise waren ähnliche Labels wie „Naturjoghurt“ und „griechischer Joghurt“ derselben Klasse zugeteilt. Die Anzahl von K=20 wurde auch deshalb als passend empfunden, da sie der Anzahl von Kategorien der BLS-Datensätze entspricht. Durch Ausprobieren, indem K bei der Evaluation der Ähnlichkeitsanalyse hoch- und heruntergesetzt wurde (K ± 1, K ± 2 und K ± 3), konnte die Entscheidung für K=20 ebenfalls bestätigt werden, da dadurch keine bessere Klassifikation erreicht wurde (die bei der Evaluation zu erwartenden Ergebnissen wurden nicht erreicht) [F].

Für eine optimale grafische Darstellung der Cluster in einem Diagramm wurde die Dimension der Nährwerte vor der Anwendung der K-Means Methode mithilfe der Hauptkomponentenanalyse (PCA, siehe Abschnitt 2.8.3) reduziert. Wie in Ng & Soo (2018) dargestellt, korrelieren Fett- und Proteinwerte von Lebensmitteln signifikant in die gleiche Richtung und in die entgegengesetzte Richtung der Kohlenhydrate [51]. Dieser Sachverhalt wird an den Abbildungen 3.10 und 3.11 ersichtlich. In den beiden Abbildungen werden die Nährwerte Kohlenhydrate zu Fett sowie Kohlenhydrate zu Proteinen im Verhältnis zueinander dargestellt. Als Datenquelle dieser Darstellungen diente die Datenbank des BLS. Durch die Hauptkomponentenanalyse konnten die Werte für Fett und für Proteine zu einem Wert (PCA Dimension 1) zusammengefasst werden. Der Wert für Kohlenhydrate bildet den zweiten Wert (PCA Dimension 2). Anhand der beiden Werte-Dimensionen PCA Dimension 1 und PCA Dimension 2

konnte nun das K-Means Verfahren durchgeführt werden. Abbildung 3.12 stellt die aus den BLS-Daten gebildeten Klassen grafisch dar.

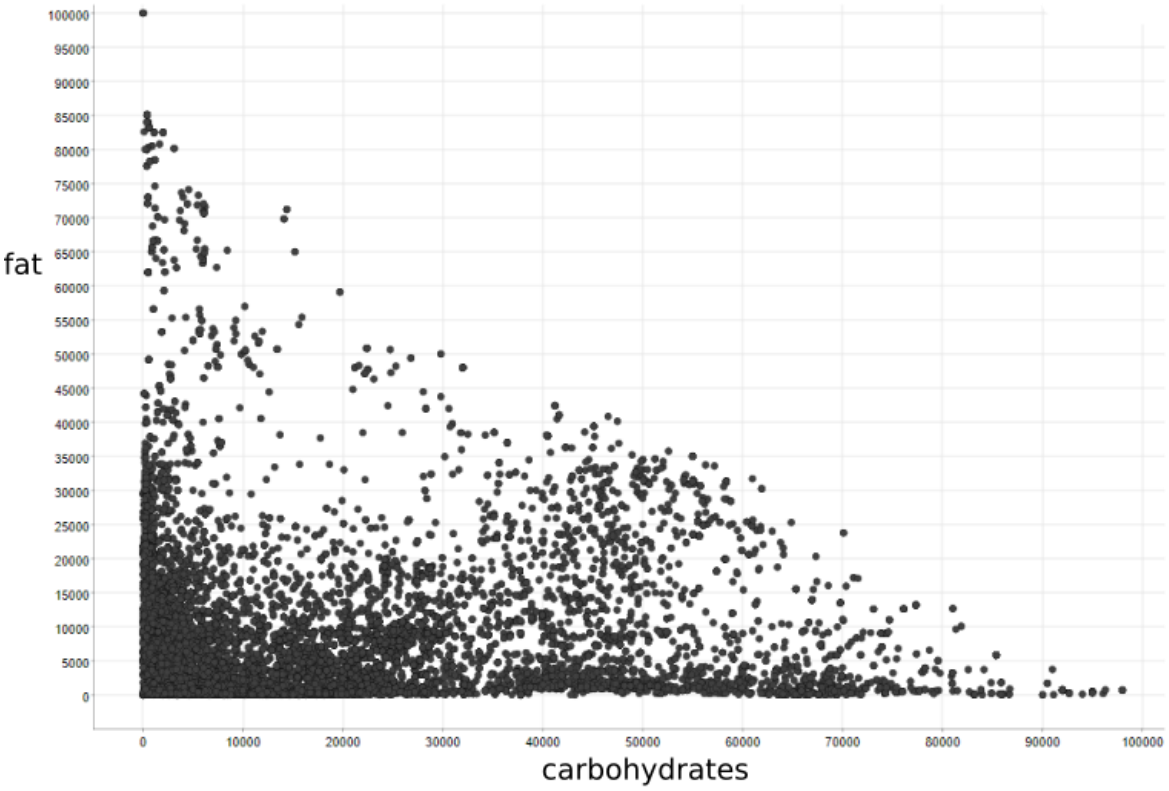


Abbildung 3.10 Darstellung des Verhältnisses von Kohlenhydraten zu Fett der BLS-Datensätze als Punktdiagramm

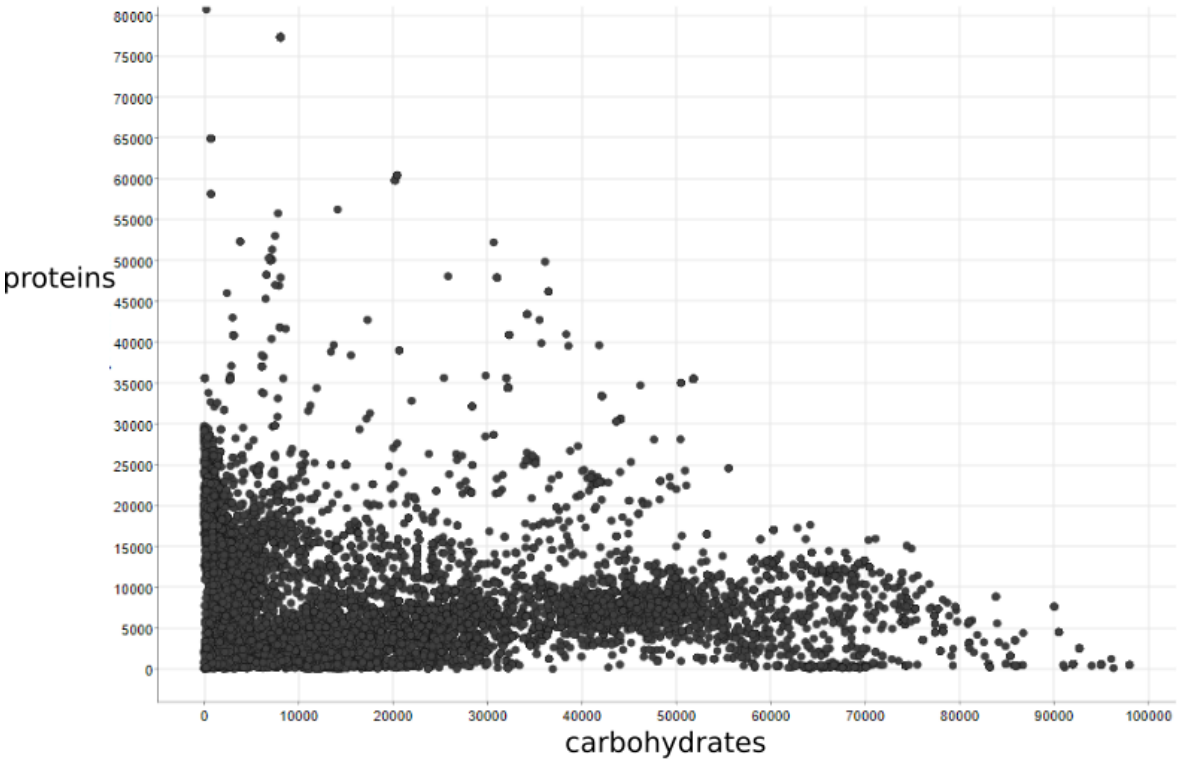


Abbildung 3.11 Darstellung des Verhältnisses von Kohlenhydraten zu Proteinen der BLS-Datensätze als Punktdiagramm

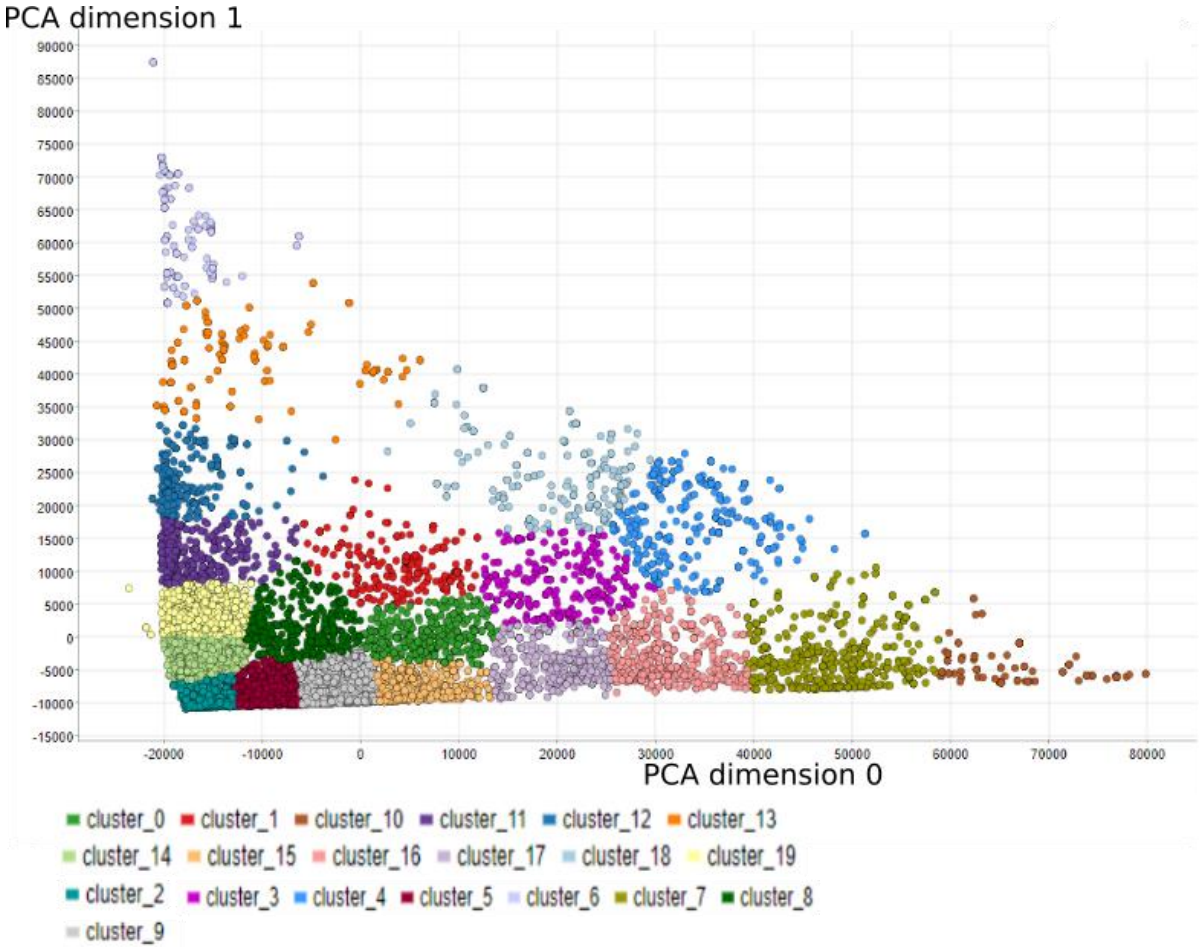


Abbildung 3.12 Darstellung der aus den BLS-Datensätzen gebildeten Clustern nach der PCA-Dimensionsreduzierung

Mithilfe der DataMining-Software KNIME werden sogenannte Workflows erzeugt, die anhand der BLS-Datensätze Wortwolken bilden (Wörter werden je nach Häufigkeit des Vorkommens in einer oder mehreren Zeichenketten unterschiedlich groß dargestellt, wobei das am häufigsten vorkommende Wort am größten dargestellt ist), welche die Häufigkeit der BLS-Kategorien der Lebensmitteldatensätze anzeigen, die den jeweiligen Klassen zugeordnet sind [F]. Beispielsweise zeigt Abbildung 3.13 die Wortwolke aller Kategorien der durch K-Means gebildeten Klasse cluster_16. Die Kategorien der Wortwolke sind in der Bedeutung ähnlich zueinander (Lebensmittel mit hohem Zuckergehalt und Schokoladenanteil). Das Bilden solcher Wortwolken bestätigte zusätzlich eine Wahl von K=20, da diese Kategorie-Bezeichner über Lebensmittelarten enthalten, deren Ähnlichkeiten plausibel begründet werden konnten (beispielsweise mehrheitlich Milchprodukte oder proteinreiche Produkte).

Abbildung 3.13 Wortwolke aus den Kategorien der BLS-Datensätze, die dem Cluster cluster_16 zugeordnet sind

3.5.3 Bestimmung der Plausibilität und Gesamtähnlichkeit

In den vorherigen Unterkapiteln wurde beschrieben, wie die Ähnlichkeitswerte für zwei Zeichenketten berechnet werden und wie die Produktdatensätze anhand ihrer Nährwerte Fett, Kohlenhydrate und Proteine in Klassen eingeteilt werden. Anhand solcher bestimmter Ähnlichkeitswerte und Klassen kann nun sowohl die Plausibilität als auch die Gesamtähnlichkeit zweier Produktdatensätze ermittelt werden. Zu einem Produktdatensatz, dessen Daten auf Plausibilität geprüft werden sollen, müssen zunächst ähnliche Produktdatensätze gefunden werden. Hierbei spielt auch die Ähnlichkeit des Produktnamens eine Rolle. Deshalb werden zunächst die Datensätze zu Produkten aus der FDWH-Datenbank selektiert, die in Teilen des Produktnamens bzw. bei denen bestimmte Wörter des Produktnamens ähnlich zu denen im zu prüfenden Produktdatensatz sind. Nun wird dieser nach und nach mit den selektierten Datensätzen auf Ähnlichkeit untersucht. Mit Hilfe von zufällig ausgewählten Produktdatensätzen werden Entscheidungsbäume generiert, die für die Plausibilitätsprüfung herangezogen werden.

Eine Entwicklung dieser Entscheidungsbäume erfordert vorab eine Bestimmung der Grenzwerte an den Knoten. Dafür werden Testdatensätze untersucht, damit festgelegt werden kann, inwiefern sich zwei Produkte ähneln müssen (welche Ähnlichkeitswerte hierfür notwendig sind), damit von der Existenz eines Produktes Y auf die Plausibilität von Produkt X geschlossen werden kann. Produkt X ist hierbei das Produkt das geprüft wird und Produkt Y wurde nach der Ähnlichkeitsanalyse (als ähnliches Produkt zu Produkt X) ermittelt. Anhand von 100 zufällig ausgewählten Testdatensätze werden die Grenzwerte der Ähnlichkeiten für die Plausibilitätsanalyse festgelegt. Durch Recherchen von Produktseiten von Herstellern oder Händlern und Bilddaten von Produktverpackungen dieser Testdatensätze, werden deren Plausibilitätswerte durch eine händische Überprüfung ermittelt. Ein algorithmusbasiertes und

automatisches Generieren der Entscheidungsbäume (mittels Gini Index) tendierte zu Overfitting der Bäume (eine Überanpassung der Grenzwerte an den Baumknoten [54][66]), was zu einer komplexen Baumstruktur führte, die zu sehr an Einzelfälle bei der Regelfindung angepasst war. Deshalb fiel die Entscheidung, die Struktur der Bäume manuell zu bestimmen. Dies war wegen der zu erwarteten geringen Zahl an Knoten, aufgrund der geringen Zahl an Ähnlichkeitsattribute und somit wegen einer erwartbaren niedrigen Komplexität der Baumstruktur, einfach durchführbar.

Zur Festlegung der Entscheidungsbäume müssen zunächst eine Grundstruktur und die Grenzwerte bestimmt werden. Die Regeln, die durch die Baumstruktur dargestellt werden, werden nach den folgenden Annahmen gebildet.

- Die Nährwerteinformationen eines Produktes X sind plausibel ($N_{plaus} = 1$), wenn ein weiteres Produkt Y existiert, dessen Nährwerte der gleichen (nach K-Means ermittelten) Klasse wie Produkt X zugeordnet sind (die Klasse muss zwingend bestimmt sein: $P_{cluster} \neq NULL$) und ein bestimmter Grenzwert für die Ähnlichkeit der Produktnamen (s_{name}) beider Produkte erreicht wird.
- Die Inhaltsstoffliste eines Produktes X ist plausibel ($I_{plaus} = 1$), wenn ein weiteres Produkt Y existiert und bestimmte Grenzwerte für die Ähnlichkeit von Produktnamen sowie Inhaltsstofflisten (s_{name} und s_{ing}) beider Produkte erreicht werden. Hierbei werden zwei Fälle betrachtet:
    1. Es wird ein bestimmter (vergleichsweise hoher) Grenzwert für s_{ing} erreicht, dafür muss ein bestimmter (vergleichsweise geringer) Grenzwert für s_{name} erreicht werden.
    2. Es wird in Bezug zum 1. Fall ein niedrigerer Grenzwert für s_{ing} erreicht. Dafür muss ein höherer Grenzwert als beim 1. Fall für s_{name} erreicht werden.

Es gilt nun, die Grenzwerte für die Ähnlichkeitswerte s_{name} und s_{ing} zu bestimmen. Die nach dem Verfahren der Kosinus-Distanz-Berechnung ermittelten Ähnlichkeitswerte sind bis auf eine Nachkommastelle gerundet. Es werden jeweils schrittweise Ähnlichkeitswerte von 0,1 bis 1,0 als Grenzwerte eingestellt und geprüft, wann der jeweilige Entscheidungsbaum die höchste Genauigkeit an richtigen Vorhersagen erzielt. Dies geschieht mit Hilfe der Berechnung des Genauigkeitswertes *ACC* (Rate von richtig als plausibel und implausibel bestimmten Testdaten im Verhältnis zu den Testdaten insgesamt, siehe Abschnitt 2.10).

Durch dieses Verfahren wurden die beiden, in Abschnitt 3.14 und 3.15 dargestellten, Entscheidungsbäume erstellt. Der während der Bestimmung der Grenzwerte am höchsten erzielte Genauigkeitswert (*ACC*), war bei der Bestimmung von N_{plaus} 0,83 (83%). Bei der Plausibilitätsbestimmung von I_{plaus} war $ACC = 0{,}87$ (87%).

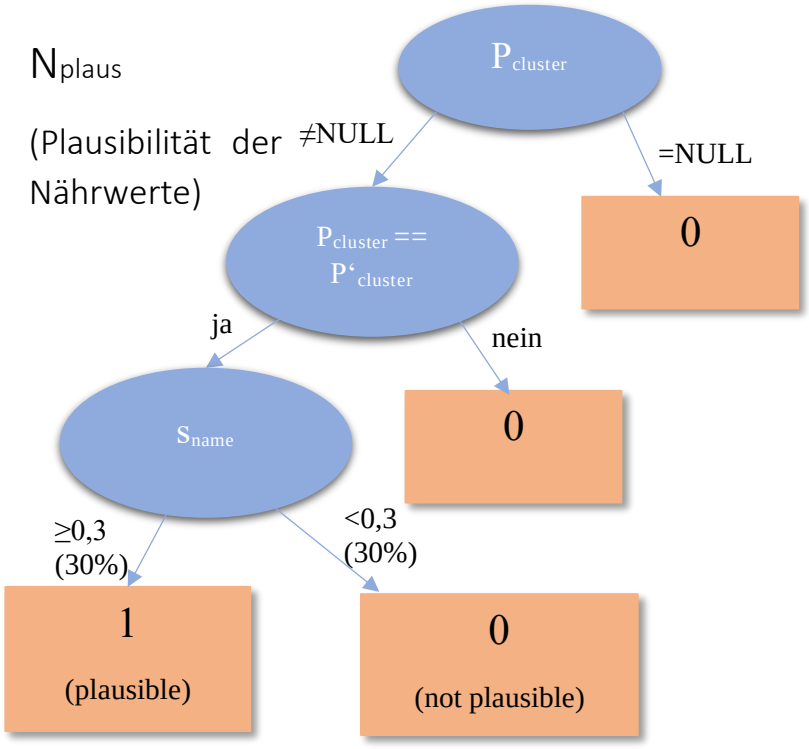


Abbildung 3.14 Entscheidungsbaum zur Bestimmung der Plausibilität der Nährwerte

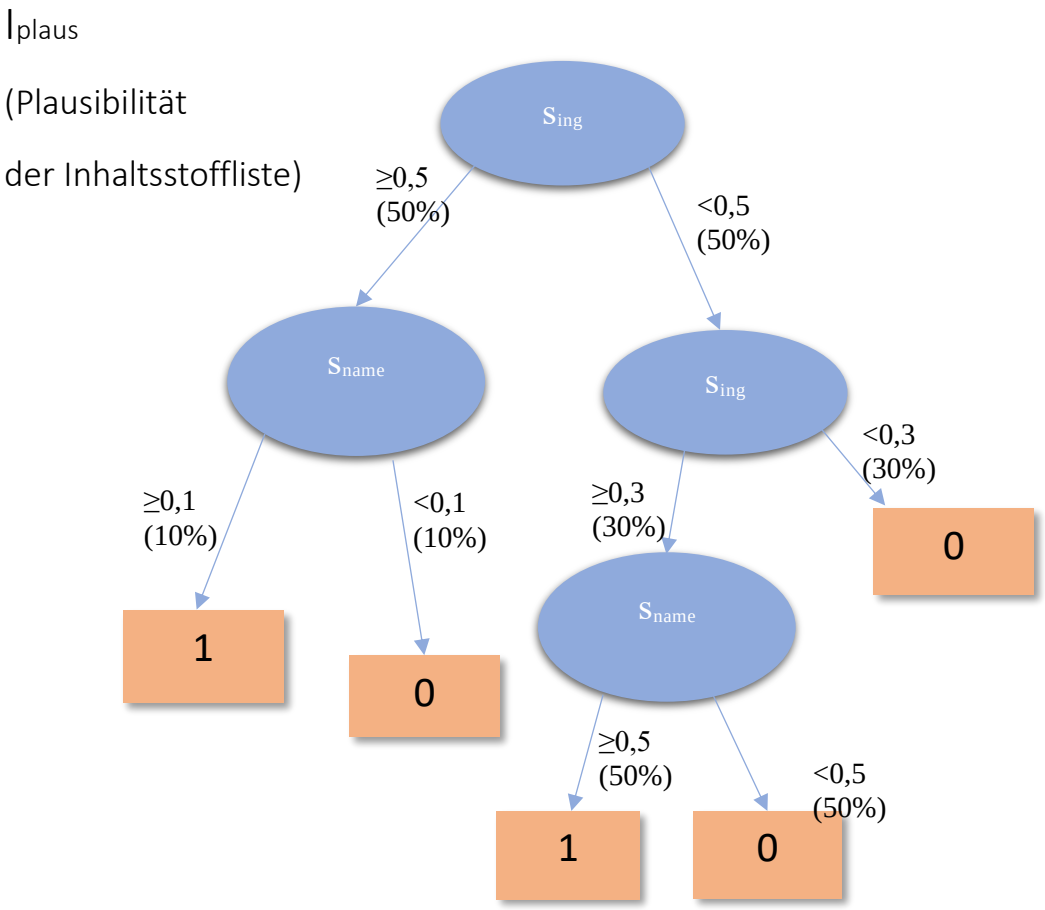


Abbildung 3.14 Entscheidungsbaum zur Bestimmung der Plausibilität der Inhaltsstoffe

Zusätzlich zu der Bestimmung der Plausibilität von Nährwertinformationen und Inhaltsstofflisten wird festgelegt, dass eine Gesamtähnlichkeit zweier zu vergleichender Produktdatensätze vorliegt, wenn folgendes gilt:

$$T_{plaus} \begin{cases} 1, \ wenn \ N_{plaus} = 1 \ und \ I_{plaus} = 1 \\ 0, \ wenn \ N_{plaus} = 0 \ oder \ I_{plaus} = 0 \end{cases} \quad (1)$$

Somit liegt eine Gesamtplausibilität (T_{plaus}) vor, wenn die beiden Produktdatensätze sowohl im Bereich der Nährwerte als auch im Bereich der Inhaltsstofflisten ähnlich sind. Die Produkte, die zu einem vom Anwender der FDWH-API abgefragten Produktdatensatz ähnlich sind, werden dem Anwender durch die API mitgegeben. Nun können weitere Produktmerkmale zweier ähnlicher Produkte verglichen und Abweichungen festgestellt werden. Mit den genannten Testdaten werden zwar die Grenzwerte an den Knoten der Entscheidungsbäume bestimmt, sie werden aber nicht zur Bewertung der Modelle

verwendet. Die Vorhersagegenauigkeit des Verfahrens wird in Abschnitt 3.8.2 anhand weiterer unabhängiger Testdaten evaluiert.

3.5.4 Entscheidung für die Anwendung der beschriebenen Methoden zur Ähnlichkeits- und Plausibilitätsanalyse

Bei der Bestimmung von Methoden des maschinellen Lernens bzw. Data Mining wurden verschiedene Verfahren getestet, die durch Recherche diesbezüglicher Literatur und bereits erprobter Ansätze ermittelt wurden. Bei der Klassifizierung von Lebensmittelproduktdaten anhand der Nährwerte für eine Ähnlichkeitsbestimmung wurde die Entscheidung getroffen, ein unüberwachtes Lernverfahren anzuwenden, bei dem keine zu lernende Klassifizierung anhand von Trainingsdaten vorgegeben wird. Eine solche Klassifizierung noch existiert nicht in den zur Verfügung gestellten Datenquellen. Vorab wurde eine Methode erprobt, bei der Lebensmittelproduktdatensätze anhand fest definierter Nährwertebereiche mittels Entscheidungsbäumen klassifiziert wurde [E]. Auch sogenannte künstliche neuronale Netze wurden zur Anwendung einer solchen Klassifizierung von Nährwertebereichen in Aussicht gestellt. Im weiteren Verlauf des Projektes wurde festgestellt, dass der Ansatz des Clustering zur Klassifizierung der Nährwerte zur Ähnlichkeitsbestimmung besser geeignet ist, da es bessere Ergebnisse lieferte und keine Auswahl solcher (ggf. ungenauen) manuellen Einstufungen in Nährwertebereiche erforderte. Außerdem wurde das gewählte Verfahren bereits in ähnlichen Anwendungsbereichen eingesetzt [L14][F].

Zur Ähnlichkeitsbestimmung zweier Zeichenketten im Zuge der Ähnlichkeitsbestimmung von Produktnamen und Inhaltsstofflisten wurde die Methode der Kosinus-Distanz-Berechnung gewählt. Vorab wurde eine Anwendung der Methode der Ähnlichkeitsbestimmung nach Jaccard in Betracht gezogen, bei der die absolute Durchschnittsmenge der Inhaltsstoffe durch die absolute Vereinigungsmenge geteilt wird [L47]). Von der Berechnung der Jaccard-Ähnlichkeit wurde abgesehen, da diese zu stark von der Anzahl der Inhaltsstoffe in der Liste abhängig ist [L18] und damit in Zwischentests schlechtere Ergebnisse bezüglich der Genauigkeitsrate (*ACC*, siehe Abschnitt 2.10) geliefert hatte, da viele gering enthaltene Inhaltsstoffe, die nur in einer von zwei zu vergleichenden Listen enthalten waren, einen großen Einfluss auf das Gesamtergebnis hatten.

Die Entscheidung zwischen plausiblen und implausiblen Daten anhand der vorab ermittelten Ähnlichkeiten wurde durch Entscheidungsbäume getroffen. Eine solche Entscheidung wäre beispielsweise auch durch die Anwendung neuronaler Netze möglich gewesen. Darauf wurde dennoch verzichtet, da eine Entscheidung mittels Entscheidungsbäume im Gegensatz zu neuronalen Netzen durch die Darstellung des Baumes besser nachzuvollziehen ist [17][L19]. Da nur wenige Werte zur diesbezüglichen Entscheidungsfindung beitrugen (nur die Ähnlichkeitswerte zweier Datensätze von

Produktnamen und Inhaltsstoffliste bzw. Nährwerteclustern), war die Performance des Entscheidungsbaumes ausreichend gut und keine Anwendung von weiteren Lösungen notwendig wie beispielsweise neuronaler Netzwerke. Diese hätten den Vorteil bei der Klassifikation anhand einer sehr großen Anzahl von Werten, da hier Entscheidungsbäume sehr unübersichtlich sind und neuronale Netze bei großen Mengen von Werten oft genauer klassifizieren [17][L19].

3.6 Konzipierung einer kontextbasierten Lebensmittel-produktdaten-API

3.6.1 Die JSON-API

Damit eine Gesundheitsapp, die auf Qualität und Plausibilität geprüften Lebensmitteldaten für ihre Zwecke empfangen kann, wurde eine entsprechende API entwickelt. Diese API unterliegt der Representational State Transfer Architektur (REST) und dient der losen Kopplung und Interoperabilität zwischen Systemen mit unterschiedlicher Implementierung [71]. Die Übermittlung von Parametern erfolgt über das JSON-Format. Ein Anwender sendet über die jeweilige App eine User-ID (*uid*) und ein Passwort-Hashwert (*pwd*) für die Authentifizierung. Als Methode (*method*) wird „getProductOverview“ für die Abfrage einer Produktübersicht gesendet. Des Weiteren wird eine Zeichenkette (*query*) mit der jeweiligen Bezeichnung oder dem Markennamen eines Produktes bzw. der Produkte, deren Informationen von der API angefragt werden sollen, und die maximale Anzahl von gewünschten Ergebnissen (*maxNum*) gesendet. Abbildung 3.16 stellt ein Beispiel einer solchen JSON-Anfrage dar.

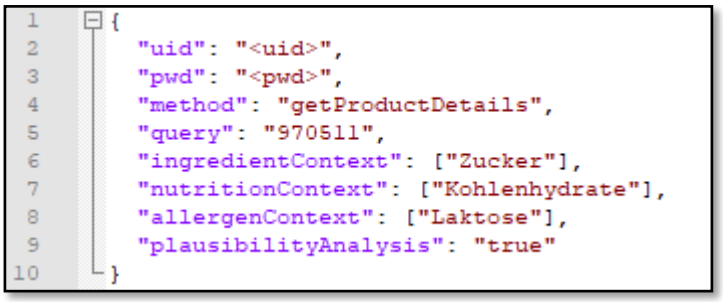
```
{
  "uid": "<uid>",
  "pwd": "<pwd>",
  "method": "getProductOverview",
  "query": "Milchreis Zimt",
  "maxNum": 100
}
```

Abbildung 3.15 Beispiel einer FDWH-API-Anfrage für eine Produktübersicht im JSON-Format

Wenn *maxNum* nicht gesetzt ist, so werden maximal die 100 ersten Ergebnisse geliefert. Enthält *maxNum* die 0 so werden unbegrenzt viele Ergebnisse gesendet, dies verlängert jedoch die Abfragezeit oft deutlich.

Als Ergebnis der Anfrage wird eine Produktübersicht mit allen, durch den oben genannten *query*-Parameter selektierten Produkte geliefert. Die Produktübersicht enthält Informationen wie Produkt-ID (entspricht in diesem Fall der Produktnummer), Produktname, Markenname, Herkunftsland, GTIN und Kategorien aller durch die API ermittelten Produkte. Im Anhang (Anhang 2) befindet sich ein Auszug eines solchen Anfrageergebnisses als Beispiel mit Bezug zu der JSON-Anfrage aus Abbildung 3.16.

Anhand von einer der soeben ermittelten Produkt-IDs kann nun eine Anfrage zur Ermittlung der Produkt-Details an die API gesendet werden. Die Produkt-Details enthalten zusätzliche Informationen des Produktes zu Inhaltsstoffliste, kennzeichnungspflichtigen Allergenen und Nährwerten. Die API wurde entsprechend der Zielsetzung der vorliegenden Arbeit so konzipiert, dass zusätzlich zu den Parametern *uid* und *pwd*, „getProductDetails“ als *method* und der Produkt-ID als *query* verschiedene Kontexte mit Informationen zu Inhaltsstoffen, Nährwerten oder kennzeichnungspflichtigen Allergenen als JSON-Parameter mitgegeben werden können. Der Vorteil dabei ist, dass hierbei der Fokus auf den jeweiligen Anwendungskontext der Gesundheitsapp gelegt werden kann, welche die Abfrage sendet. Möchte ein Anwender beispielsweise wissen, ob in dem Produkt Zucker und Kohlenhydrate sowie Laktose enthalten ist (im speziellen Kontext des Anwendungszwecks in diesem Beispiel sind diese Informationen von besonderer Wichtigkeit), so gibt dieser die genannten Informationen an die API weiter. In dem vorliegenden Beispiel würde er als Parameter *ingredientContext* den Wert „Zucker“, als Parameter *nutritionContext* den Wert „Kohlenhydrate“ und als *allergenContext* den Wert „Laktose“ übergeben. Mit dem Parameter *plausibilityAnalysis* wird mit dem Wert „true“ mitgeteilt, dass die Ähnlichkeits- und Plausibilitätsanalyse durchgeführt werden soll (siehe Abbildung 3.17).



```
{
  "uid": "<uid>",
  "pwd": "<pwd>",
  "method": "getProductDetails",
  "query": "970511",
  "ingredientContext": ["Zucker"],
  "nutritionContext": ["Kohlenhydrate"],
  "allergenContext": ["Laktose"],
  "plausibilityAnalysis": "true"
}
```


Abbildung 3.16 Beispiel einer FDWH-API-Anfrage für Details eines Produktes im JSON-Format

Die Antwort der Anfrage aus Abbildung 3.17 mit den Produktdetails befindet sich ebenfalls im Anhang (Anhang 3). Als Informationen zu den mitgegebenen Kontexten werden jeweils die Wahrheitswerte „wahr“ (true) mitgegeben, falls die jeweilige angefragte Information für das Produkt zutrifft oder „falsch“ (false) falls sie nicht zutrifft [F].

3.6.2 Grafische Benutzerschnittstelle

Zur besseren Darstellung und Vereinfachung von Tests während der Evaluation wurde eine grafische Benutzerschnittstelle für API-Anfragen als Weboberfläche implementiert. Somit lassen sich das Senden der *query*, das Anzeigen der Ergebnisse als Produktübersicht, die Übergabe der Kontext-Parameter und das Anzeigen der Produktdetails durch die grafische Weboberfläche steuern [F]. Die verschiedenen Fenster dieser Weboberfläche werden in Abbildung 3.18 nach der Reihenfolge des Aufrufs nummeriert dargestellt.

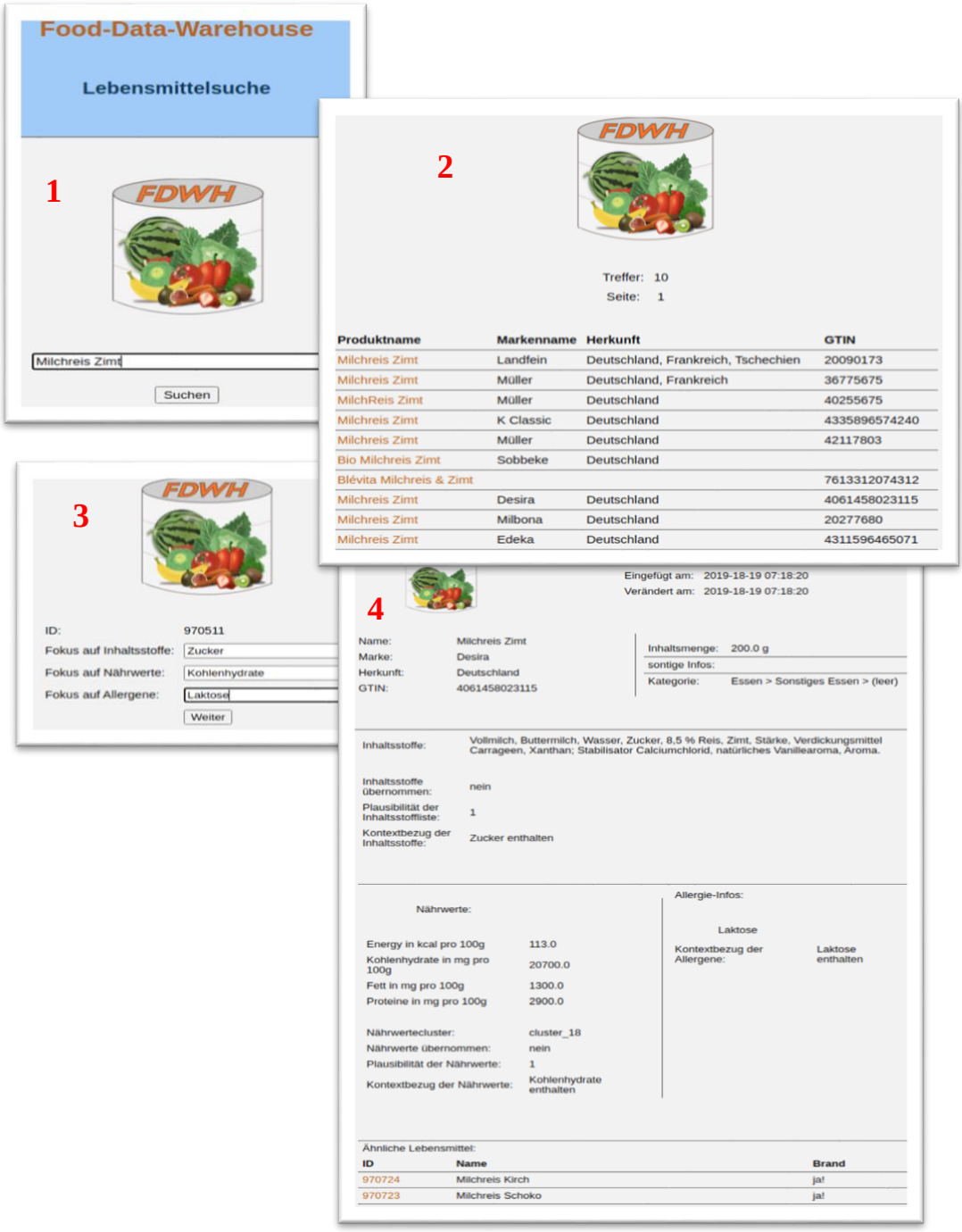


Abbildung 3.17 Weboberfläche für Testabfragen innerhalb des FDWH

3.6.3 Die Suchfunktionen des FDWH

Für die Lebensmittelproduktsuche im FDWH wurden spezifische Suchfunktionen implementiert, die es erlauben, die gewünschten Datensätze im FDWH über die API zu selektieren und im JSON-Format auszugeben. Die Produktsuche (bzw. Produktselektierung in der FDWH-Datenbank) wird anhand der nachfolgend beschriebenen Regeln durchgeführt. Der jeweiligen Suchfunktion wird die Suchzeichenkette (*query*) und die maximale Anzahl an möglichen Ergebnissen (*maxNum*) mitgegeben. Alle Sonderzeichen in *query* werden durch Leerzeichen ersetzt. Bei der Suche wird die Groß- und Kleinschreibung ignoriert.

1. Wenn *query* einer GTIN entspricht, selektiere die Produktdatensätze anhand der GTIN. Die Suche ist nach diesem Schritt beendet.
2. Selektiere alle Produkte, deren Namen exakt *query* entsprechen.
   - Beispiel: *query* enthält „Coca Cola Light“ und Name von selektiertem Produkt enthält „Coca Cola Light“.
3. Selektiere alle Produkte, bei denen *query* einem Teil des Produktnamens entspricht, der von dem restlichen Teil durch Leer- oder Sonderzeichen getrennt ist.
   - Beispiel: *query* enthält „Coca Cola Light“ und Name von selektiertem Produkt enthält „Coca Cola Light ohne Zucker“.
4. Selektiere alle Produkte, bei denen *query* einem Teil des Produktnamens entspricht.
   - Beispiel: *query* enthält „Coca Cola Light“ und Name von selektiertem Produkt enthält „Coca ColaLight“.
5. Teile *query* in alle Kombinationen zweier Wortpaare auf. Selektiere alle Produkte, bei denen eine der Kombinationen einem Teil des Produktnamens entspricht, der von dem restlichen Teil durch Leer- oder Sonderzeichen getrennt ist.
   - Beispiel: *query* enthält „Coca Cola Light“. Kombinationen sind: „Coca Cola“, „Cola Light“, „Coca Light“. Selektierte Produkte wären unter anderem „Coca Cola“, „Coca Cola Zero“, Pepsi Cola Light“.
6. Teile *query* in alle Kombinationen zweier Wortpaare auf. Selektiere alle Produkte, bei denen eine der Kombinationen einem Teil des Produktnamens entspricht, der von dem restlichen Teil durch Leer- oder Sonderzeichen getrennt ist.
   - Beispiel: *query* enthält „Coca Cola Light“. Kombinationen sind: „Coca Cola“, „Cola Light“, „Coca Light“. Selektierte Produkte wären unter anderem „Coca ColaZero“, CocaCola Light“.
7. Teile *query* in einzelne Wörter auf. Selektiere alle Produkte, bei denen eines der Wörter einem Teil des Produktnamens entspricht, der von dem restlichen Teil durch Leer- oder Sonderzeichen getrennt ist.

- Beispiel: *query* enthält „Coca Cola Light“. Suchwörter sind: „Coca“, „Cola“, „Light“. Selektierte Produkte wären unter anderem „Coca Cola“, „Pepsi Cola“, Gouda Light“.

8. Teile *query* in einzelne Wörter auf. Selektiere alle Produkte, bei denen eines der Wörter einem Teil des Produktnamens entspricht.
   - Beispiel: *query* enthält „Coca Cola Light“. Suchwörter sind: „Coca“, „Cola“, „Light“. Selektierte Produkte wären unter anderem „CocaCola“, „Chocolate“.
9. Selektiere Produkte anhand ihrem Sound-Expression-Code (SOUNDEX-Code). Hier wird geprüft, ob der SOUNDEX-Code der *query* dem SOUNDEX-Code eines Produktes im FDWH entspricht.
10. Selektiere Produkte anhand ihres Markennamens anstelle des Produktnamens. Wiederhole hierbei die Schritte 2 bis 9.
11. Selektiere Produkte anhand ihres Kategorienamens anstelle des Produktnamens. Wiederhole hierbei die Schritte 2 bis 9.

Die aufgelisteten Suchfunktionen werden nacheinander der Reihe nach abgerufen. Somit werden Ergebnisse in der Reihenfolge ausgegeben, wie sie durch die aufgerufenen Regeln selektiert werden bis *maxNum* erreicht ist. Dies soll dazu führen, dass die Produktergebnisse nach Genauigkeit sortiert aufgelistet sind, in Bezug zu der *query*. In Schritt 9 wird anhand von SOUNDEX-Codes selektiert, um gegebenenfalls auch bei Suchworten mit Rechtschreibfehlern die richtigen Produkte zu finden. Die SOUNDEX-Funktion codiert eine Zeichenkette phonetisch. Ähnlich klingende Buchstaben enthalten die gleiche Nummerierung im Code [55]. Ein Teil der Suchfunktionen stammt aus der im DiDiER-Projekt entwickelten API. Dieser Teil wurde weiterentwickelt und für das FDWH wiederverwendet.

3.7 Performante Bearbeitung großer Datensätze mit Hilfe von Big Data Technologie

3.7.1 Entscheidung für die Verwendung von TimescaleDB als Datenbanksystem für FDWH-Datensätze

Als Datenbank für die performante Speicherung zeitbezogener Datensätze wurden zunächst verschiedene Lösungen betrachtet. Im Gegensatz zu herkömmlichen relationalen Datenbanken, welche die Abfragesprache Structured Query Language (SQL) einsetzen, eignen sich NOSQL Lösungen besser für die Speicherung sehr vieler, bzw. sehr großer Datensätze. Dies basiert meistens auf der Möglichkeit erhöhter Zugriffsgeschwindigkeiten durch optimierte Partitionierung und Indizierung der Datensätze

sowie der Skalierbarkeit solcher Speicherlösungen auf mehrere Rechnerknoten (Cluster) [53]. Um die zeitlichen Veränderungen von Lebensmitteldatensätzen zu speichern, soll die im Projekt der vorliegenden Arbeit angewandte Datenbank ein schnelles Laden und Selektieren von zeitbezogenen Datensätzen ermöglichen. Hierzu eigenen sich am besten sogenannte Zeitreihendatenbanken wie InfluxDB, TimescaleDB, OpenTSDB und Graphite [L48]. Eine weitere betrachtete Lösung für den Einsatz mit Zeitreihendaten ist VictoriaMetrics, wobei sich herausstellte, dass sich diese Datenbank ausschließlich für die Speicherung aktueller Datensätze eignet (eine Speicherung von Daten, deren Zeitstempel in die Vergangenheit reichen, ist nicht möglich) [L61]. Aus diesem Grund wurde VictoriaMetrics nicht in Betracht gezogen, da die Lebensmittel- und Lebensmittelproduktdaten der verschiedenen Datenquellen vergangene Zeitstemple enthalten. Des Weiteren wurde ermittelt, dass sich die Datenbank Graphite ausschließlich für die Speicherung numerischer bzw. metrischer Zeitreihendaten eignet [L17][L48]. Für das Laden von Lebensmittelproduktdaten in ein Datenbanksystem ist es notwendig, dass das System die Speicherung von Zeichenketten (z. B. die Speicherung von Inhaltsstoff-Bezeichnern) unterstützt. Nach einer Analyse von Schneider verbraucht TimescaleDB im Gegensatz zu der Zeitreihendatenbank weniger Arbeitsspeicher, im Gegenzug dazu verbraucht InfluxDB weniger Festplattenspreicher [L44]. Da aber TimescaleDB die bessere Datenstrukturierung und die Möglichkeit der Speicherung komplexer Daten bietet, insbesondere wenn mehrere Metainformationen einbezogen werden, müssen bei dessen Verwendung keine zusätzlichen Datenbanken für die Metadatenspeicherung angelegt werden. InfluxDB eignet sich eher für die Speicherung metrischer Daten [L44]. Die beiden Speicherlösungen sind im Bereich der zeitlichen Abfrageperformanz in etwa gleichauf [33]. Nach Imru (2020) liefert OpenTSDB eine hohe zeitliche Performanz bei Schreibvorgängen, die Einführung in die Technologie wird aber aufgrund der komplexen Interaktionsmöglichkeiten und einer hohen Lernkurve als schwierig angesehen [L22]. OpenTSDB bietet wie InfluxDB keine nativen SQL-Abfragemöglichkeiten und einigt sich ebenso, aufgrund der eingeschränkten Möglichkeit zur Datenstrukturierung, mehr für die Speicherung großer metrischer Datensätze [L44].

Letztendlich fiel die Entscheidung auf TimescaleDB als Datenbanksystem. Der größte Vorteil dabei war, dass das System auf PostgreSQL aufbaut und aus diesem Grund die Technologien relationaler Datenbanken (unter anderem SQL) als einzige der vorgestellten Datenbanken nativ unterstützt [L55][L48]. Dies erlaubt die Verbindung zwischen Daten-Entitäten über Relationen unter der Vermeidung von Redundanzen [48] und die optimierte Strukturierung der Lebensmittel- und Lebensmittelproduktdaten sowie der Metadaten. Des Weiteren wurde dem SQL-erfahrenen Entwickler des FDWH eine zügigere Entwicklung, im Vergleich zu den restlichen genannten Datenbanksystemen ermöglicht. Mit PostgreSQL als Basissystem profitiert TimescaleDB von dessen Zuverlässigkeit und von entwickelten Tools und Plugins einer großen Community [L37][L48]. Die Datenbankstruktur des FDWH lässt sich durch die Bildung von Relationen im Snowflake-Schema abbilden, was den Vorteil

von unkomplizierten Abfragen durch die Reduzierung von Verzweigungen mit sich bringt. Wie in Abschnitt 2.9.1 beschrieben, bietet TimescaleDB schon auf einem einzigen Rechner (Single-Cluster-System) eine hohe zeitliche Schreibperformanz und Speichereinsparungen.

3.7.2 Verwendung des Apache Spark Framework in Python für die performante Datenverarbeitung

Für die Datenqualitätsoptimierung anhand der Ähnlichkeits- und Plausibilitätsüberprüfung, wird das in Abschnitt 2.9.2 vorgestellte Framework PySpark verwendet. Dies ermöglicht die eine Plausibilitätsprüfung von Lebensmittelproduktdaten direkt nach deren Anfrage durch die API des FDWH (siehe Abbildung 3.19). Dies ist von Vorteil, da neue Lebensmitteldatensätze jederzeit zum FDWH hinzugefügt werden können und ansonsten immer nach einem solchen Einfügen alle Daten auf Plausibilität überprüft werden müssten. Diese Art der Plausibilitätsüberprüfung wäre sehr rechenintensiv. Durch die Plausibilitätsprüfung der Daten, die gerade benötigt werden, werden somit Ressourcen eingespart. Die zur Analyse selektierten Daten des FDWH werden als RDDs geladen. Deren Ergebnisse werden zwischengespeichert und können bei nachfolgenden Abfragen wiederverwendet werden. Die RDDs können durch das PySpark-Framework auch als Dataframes dargestellt werden. Dataframes sind virtuelle Datentabellen ähnlich denen des Pandas Framework auf die implementierten Operationen angewendet werden.

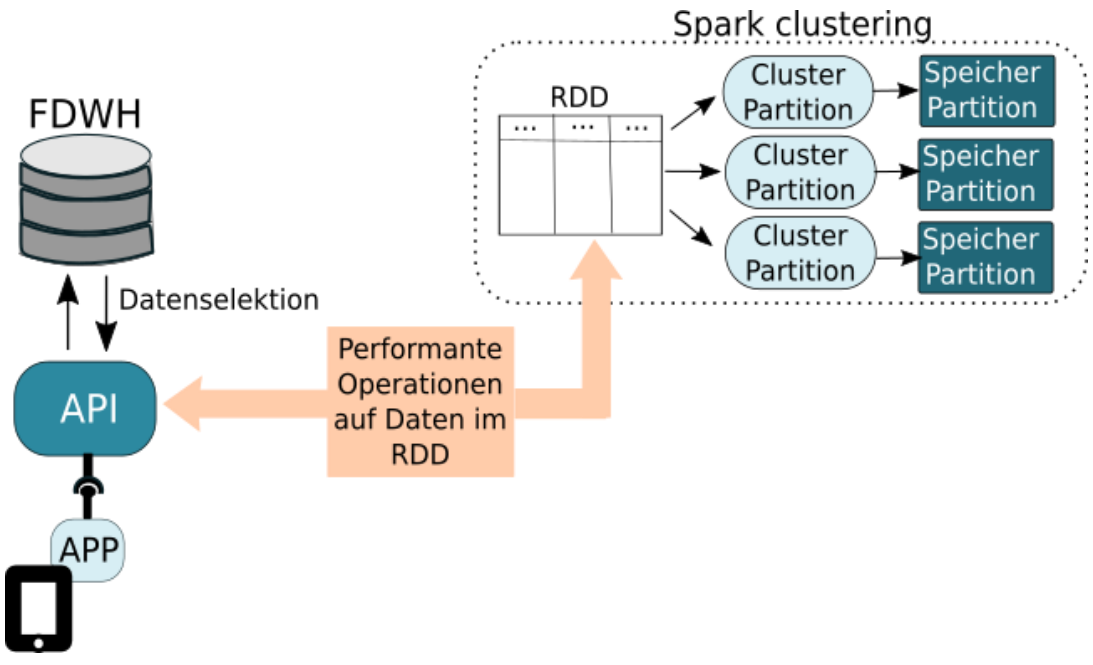


Abbildung 3.18 Darstellung der FDWH-API-Anbindung an das Spark Framework und dessen Bearbeitung der Daten im Hauptspeicher sowie an die App des Anwenders

3.8 Auswertung und Evaluierung

3.8.1 Auswertung der durch Data Profiling bereinigten Attributwerte

Durch den Data Profiling Prozess werden Werte der verschiedenen Attribute des FDWH die nach Extrahierung aus den externen Datenquellen nicht in das jeweilige standardisierte Format des FDWH transformiert werden können (beispielsweise nicht in einen bestimmten Datentyp konvertiert werden können), aussortiert. Des Weiteren werden Werte aussortiert, die während der Transformation nicht den Regeln der regulären Ausrücke und des Data Profiling entsprechen. Durch einen Test, bei dem die Daten aller verwendeten externen Datenquellen in die zentrale FDWH-Datenbank geladen wurden (Stand: 18.05.2021) wurden nach dem Vergleich, welche Daten jeweils durch Data Profiling aussortiert wurden, die Zahlen in Tabelle 3.6 ermittelt. Diese Tabelle zeigt die Anzahl der jeweiligen Attributwerte, die nach der Anwendung des Data Profiling ins FDWH geladen wurden. Des Weiteren ist die Gesamtanzahl der jeweiligen Informationen aller Datenquellen vor und nach dem Data Profiling Prozess in der Tabelle aufgelistet.

Attribute	WikiFood	OpenFood Facts	Food Repo	das-ist-drin	BLS	Danone	total (after cleansing)	before cleansing
food name	66884	54971	14892	996	14814	94	152651	176588
brand	61388	56592	0	996	0	94	119070	131792
origin	51102	63818	14762	0	0	94	129682	160778
gtin	55350	61995	29649	983	0	0	147977	161680
additional information	8102	45343	0	996	0	94	54535	54535
content	57749	49916	29982	0	0	94	137741	160778
allergen information	17038	14517	12343	465	0	93	44.486	17503
ingredient list	50151	35651	28329	988	0	94	115213	161774
nutrition fact	33863	35302	29015	996	14814	94	114084	176588
entire dataset	66884	63818	29982	996	14814	94	176588	176588

Tabelle 3.6 Anzahl der Attributwerte der jeweiligen Datenquellen insgesamt, vor und nach der Bereinigung

Abbildung 3.20 stellt die Anzahl der Gesamtattributwerte jeweils vor und nach dem Data Profiling und anschließendem Bereinigungsprozess grafisch dar.

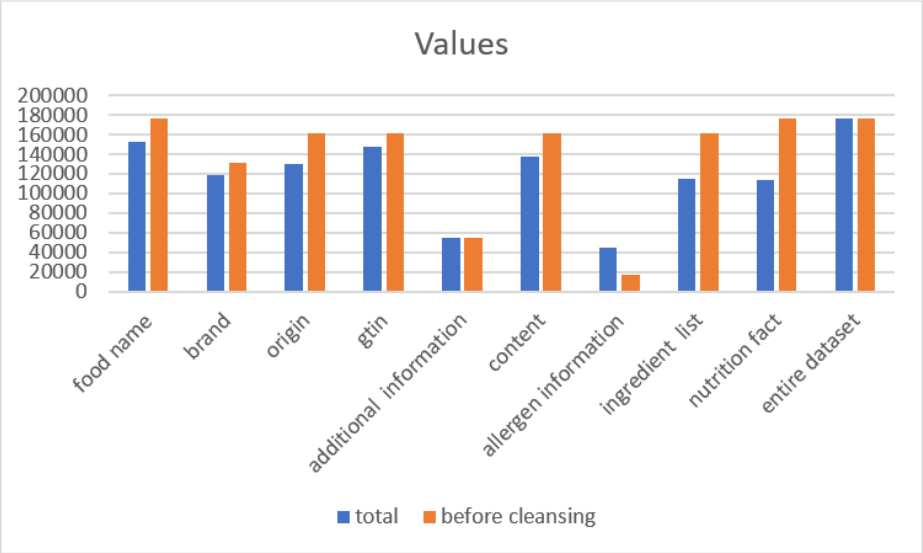


Abbildung 3.20 Anzahl der Gesamtattributwerte vor und nach der Bereinigung

Wie in Abbildung 3.20 zu erkennen ist, sind die Werte bzgl. Inhaltsstoffliste und Nährwert-informationen nach dem Data Profiling besonders zurückgegangen. Dies liegt mitunter daran, dass für diese Werte in einigen der Datenquellen Platzhalter für fehlende Werte eingesetzt wurden, die den Regeln der oben genannten regulären Ausdrücke nicht entsprachen (beispielsweise die Zeichenkette „-#-“ für fehlende Inhaltsstofflisten in den OFF-Datenquellen oder negative numerische Wert für fehlende Nährwerte). Außerdem konnten viele dieser fehlenden Werte nicht durch Data Profiling Regeln ermittelt werden, da notwendige Informationen fehlten. Z. B. fehlten in vielen der Datensätze die Nährwertinformationen komplett und es waren keine Prozentangaben in den Inhaltsstofflisten vorhanden, wodurch aufgrund der fehlenden Verknüpfung keine Berechnung der Werte durchgeführt werden konnte. Auch Werte wie Verpackungsinformationen (unter dem Attribut additional_info gespeichert) oder Informationen über kennzeichnungspflichtige Allergene (unter dem Attribut allergen_info gespeichert) werden nicht durch jede der externen Datenquellen geliefert und sind somit insgesamt am wenigsten im FDWH enthalten. Hierbei können im Falle der Verpackungsinformationen keine Regeln zur Vervollständigung abgeleitet werden, aufgrund fehlender Verknüpfungen zu weiteren Attributen. Im Falle der kennzeichnungspflichtigen Allergene konnten einige Datensätze durch die Verknüpfung mit Informationen der Inhaltsstoffliste vervollständigt werden.

Letztendlich lässt sich zusammenfassen, dass die vorgestellten Data Profiling Regeln Methoden bei syntaktischen Sachverhalten effektiv sind, beispielsweise der Erkennung und Umwandlung von Zeichenketten nicht der festgelegten Struktur entsprechen (z. B. mit Hilfe von regulären Ausdrücken), da hier feste Regeln definiert sind. Wie Tabelle 3.20 zeigt, wurden einige Werte nach der Bereinigung durch den Data Profiling Prozess aussortiert. Aussortierte Werte sind solche, für die es nicht gelingt sie (anhand fester Regeln) in eine konsistente Struktur umzuwandeln bzw. den korrekten Wert zu bestimmen. Bei Informationen in Datensätzen bei denen Fehler nicht auf der syntaktischen Struktur, sondern der semantischen Bedeutung beruhen, gestaltet sich die Fehlererkennung und Korrektur schwieriger. Hier können lediglich feste Regeln zur Bestimmung oder Berechnung anhand der Informationen in Datenattributen und Metadaten behilflich sein. Ein Einsatz des maschinellen Lernens bzw. Data Mining hilft bei der Ermittlung von Informationen aus Datensätzen untereinander. Ein solcher Ansatz wurde für die Ähnlichkeits- und Plausibilitätsanalyse gewählt, dessen Evaluierung im nachfolgenden Abschnitt stattfindet.

3.8.2 Evaluierung und Auswertung der Ähnlichkeits- und Plausibilitätsanalyse

3.8.2.1 Selektion der Daten

Zur Auswertung der Ähnlichkeits- und Plausibilitätsanalyse wurden Daten aus der DiDiER-Studie verwendet. Während der Studie führten 18 Probanden ein digitales Ernährungstagebuch, in dem Informationen von verzehrten Lebensmitteln festgehalten wurden. Für die Auswertung wurden Lebensmitteldaten im FDWH selektiert und ausgewertet. Zunächst war eine Auswertung der Daten von drei zufällig ausgewählten Tagebüchern (insgesamt 123 Lebensmitteldatensätze) vorgesehen. Während der Selektierung ist aufgefallen, dass überwiegend Daten gleicher Hersteller mit Bio-Siegeln [L2] und Daten von nicht zusammengesetzten Lebensmitteln (ohne Inhaltsstoffliste) in den ausgewählten Tagebüchern vorhanden waren. Aufgrund der Gefahr einer diesbezüglichen Stichprobenverzerrung wurden zusätzlich 100 Testdaten aus dem FDWH zur Auswertung selektiert. Die Auswahl der Testdaten fand zufällig und im Verhältnis zu ihrer Anzahl in den jeweiligen Datenquellen statt. Letztlich wurden 40 OpenFoodFacts.org-Datensätze, 40 WikiFood.eu-Datensätze, 19 FoodRepo.org-Datensätze und ein das-ist-drin.de-Datensatz aus dem FDWH selektiert. Bundeslebensmittelschlüssel-Datensätze wurden nicht ausgewählt, da wegen der Prüfung durch das MRI eine hohe Datenqualität angenommen wird und keine Inhaltsstofflisten in den Lebensmitteldatensätzen enthalten sind. Es fiel ebenso die Entscheidung, dass in diesem Hinblick auf die Selektierung von Danone-Daten verzichtet wird, wegen der geringen Anzahl dieser Datensätze und der hohen Datenqualität, da diese direkt vom Hersteller stammen. Aufgrund der genannten Sachverhalte von BLS- und Danone-Daten könnte es bei einer Verwendung bei der Auswertung ebenfalls zu Stichprobenverzerrungen kommen.

Zu ergänzen ist hier, dass im Bereich der Inhaltsstoffattribute die selektierten Tagebuch- und Testdatensätze bei der Ähnlichkeits- und Plausibilitätsanalyse ausgenommen wurden, die nicht zusammengesetzte Lebensmittel repräsentierten. Diese Datensätze enthalten keine Inhaltsstofflisten, bzw. deren Inhaltsstofflisten enthalten nur eine einzige Zutat, nämlich das Lebensmittel selbst. Diese Datensätze können im Bereich der Inhaltsstoff-Attributwerte nicht im vernünftigen Maße auf Plausibilität überprüft werden. Deshalb reduziert sich die Anzahl der verwendeten Testdatensätze geringfügig. Durch eine händische Überprüfung der Tagebuch- und Testdatensätze auf Produktseiten von Herstellern oder Händlern und Bilddaten von Produktverpackungen, werden deren tatsächlichen Plausibilitätswerte (für die Eintragung in die Wahrheitsmatrizen der ROC-Analyse) ermittelt.

3.8.2.2 Berechnung von Genauigkeit, Präzision und Receiver Operating Characteristics

Sowohl die selektierten Tagebuchdatensätze als auch die selektierten Testdatensätze wurden anhand der Ähnlichkeits- und Plausibilitätsanalyse auf ihre Plausibilität in den Attributbereichen der Inhaltsstoffe und der Nährwerte untersucht. Anschließend fand die Bestimmung der Gesamtplausibilität der Testdaten statt. Wahrheitsmatrizen dienten hierbei zur Durchführung der ROC-Analyse, wie in Abschnitt 2.10 beschrieben. Die Wahrheitsmatrizen wurden für die Plausibilität der Tagebuchdatensätze im Bereich der Nährwerteattribute (DN), im Bereich der Inhaltsstoffattribute (DI) und für die Gesamtplausibilität der Tagebuchdatensätze (DG) gebildet. Ebenso wurden die Wahrheitsmatrizen für die Plausibilität der weiteren 100 Testdaten im Bereich der Nährwerteattribute (TN), der Inhaltsstoffattribute (TI) und für die Gesamtplausibilität (TG) erstellt. Die Wahrheitsmatrizen enthalten als Werte sowohl die Anzahl der durch die Ähnlichkeits- und Plausibilitätsanalyse als plausibel eingestuften Attributwerte, die tatsächlich plausibel sind (RP), als auch die Attributwerte, die als plausibel eingestuft, aber tatsächlich implausibel sind (FP). Des Weiteren sind die als implausibel eingestuften Attributwerte, die tatsächlich plausibel sind (FN) und die als implausibel eingestuften Attributwerte, die tatsächlich implausibel sind (RN) enthalten. Diese Wahrheitsmatrizen werden durch die Tabellen 3.5 bis 3.12 dargestellt.

DN		**Tatsächlich plausibel**	
		Plausibel	Implausibel
Plausibilitätstest	Positiv	68 (RP)	11 (FP)
	Negativ	7 (FN)	37 (RN)

Tabelle 3.7 Wahrheitsmatrix der Plausibilitätsanalyse der Tagebuchdatensätze im Bereich der Nährwerteattribute

DI		**Tatsächlich plausibel**	
		Plausibel	Implausibel
Plausibilitätstest	Positiv	50	9
	Negativ	11	21

Tabelle 3.8 Wahrheitsmatrix der Plausibilitätsanalyse der Tagebuchdatensätze im Bereich der Inhaltsstoffattribute

DG		**Tatsächlich plausibel**	
		Plausibel	Implausibel
Plausibilitätstest	Positiv	36	6
	Negativ	6	28

Tabelle 3.9 Wahrheitsmatrix der Gesamtplausibilitätsanalyse der Tagebuchdatensätze

TN		**Tatsächlich plausibel**	
		Plausibel	Implausibel
Plausibilitätstest	Positiv	54	9
	Negativ	11	26

Tabelle 3.10 Wahrheitsmatrix der Plausibilitätsanalyse der 100 Testdatensätze im Bereich der Nährwerteattribute

TI		**Tatsächlich plausibel**	
		Plausibel	Implausibel
Plausibilitätstest	Positiv	63	3
	Negativ	7	23

Tabelle 3.11 Wahrheitsmatrix der Plausibilitätsanalyse der 100 Testdatensätze im Bereich der Inhaltsstoffattribute

TG		**Tatsächlich plausibel**	
		Plausibel	implausibel
Plausibilitätstest	Positiv	40	6
	Negativ	13	37

Tabelle 3.12 Wahrheitsmatrix der Gesamtplausibilitätsanalyse der 100 Testdatensätze

Anhand der Werte in den Wahrheitsmatrizen wurden die Werte für RPR, FPR, PREC und ACC berechnet. Diese sind nachfolgend aufgelistet.

- DN:
  - RPR = 0,91; FPR = 0,23; PREC = 0,86; ACC = 0,85
- DI:
  - RPR = 0,93; FPR = 0,05; PREC = 0,98; ACC = 0,78
- DG:
  - RPR = 0,86; FPR = 0,18; PREC = 0,86; ACC = 0,84
- TN:
  - RPR = 0,83; FPR = 0,26; PREC = 0,86; ACC = 0,80
- TI:
  - RPR = 0,90; FPR = 0,12, PREC = 0,96, ACC = 0,90
- TG:
  - RPR = 0,76; FPR = 0,14; PREC = 0,87; ACC = 0,80

Der Wert ACC steht für die Genauigkeit der Vorhersage und liegt zwischen 0,78 und 90 (zwischen 78 % und 90 %) und somit im glaubwürdigen Bereich. Der PREC-Wert steht für die Präzision der Analyse. Dieser Wert liegt zwischen 0,86 und 0,98 (86 % und 98 %) und sagt somit ein positives Maß an Präzision voraus. Weiterhin wurden die RPR- und FPR-Werte in das ROC-Diagramm in Abbildung 3.20 eingetragen.

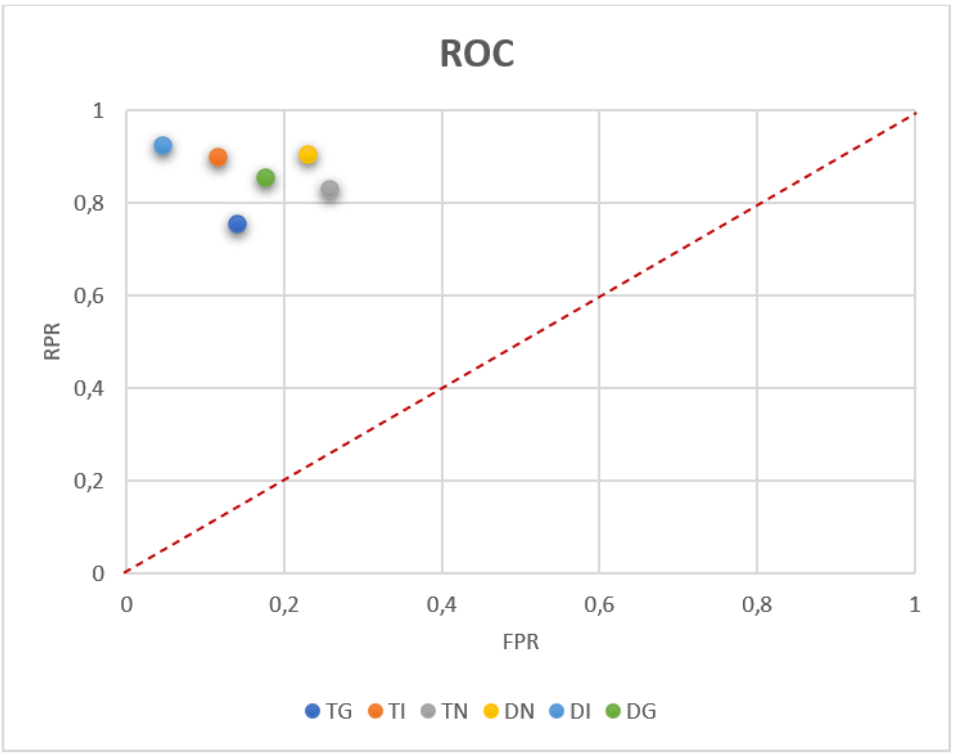


Abbildung 3.21 ROC-Diagramm mit RPR- und FPR-Daten der Validierung die sich links über der Diagonalen befinden

Wie zu sehen ist, liegen die Punkte aller Messungen im linken oberen Bereich des Diagramms nahe dem Wert (0,1) und über der Diagonalen des ROC-Graphen. Dies übermittelt uns die Erkenntnis einer guten Plausibilitätsklassifizierung durch die Ähnlichkeits- und Plausibilitätsanalyse von den Tagebuch- und Testdaten im Bereich der Attributwerte für Inhaltsstoffe und Nährwerte sowie der Gesamtplausibilität.

4 Diskussion

4.1 Interpretation der Ergebnisse und Ausblick des Produktiveinsatzes der FDWH-Daten und entwickelten Methoden

4.1.1 Datenqualitätsverbesserung durch Data Profiling sowie Ähnlichkeits- und Plausibilitätsanalyse

Die Evaluierung belegt, dass die gewählten Methoden des Data Profiling das automatisierte Erkennen von inkonsistenter und fehlender oder doppelter Datensätze erlauben. Im FDWH werden dadurch nur noch Werte gespeichert, die den Regeln in Abschnitt 3.2 entsprechen, die durch das Data Profiling generiert wurden. Wie die Auswertungen in Abschnitt 3.8.1 zeigen, reduzierte sich die Anzahl der meisten im FDWH vorhandenen Attributwerte deutlich im Vergleich zu den Datenquellen, aus denen die Werte extrahiert wurden. Insgesamt ist erkennbar, dass durch die beschriebenen Regeln die Einhaltung einheitlicher Formate und Datentypen und eine konsistente Struktur der Daten im FDWH sichergestellt werden. Fehlende Attributwerte werden erkannt und können in einigen Fällen anschließend berechnet oder anderweitig ermittelt werden. Insgesamt werden fehlerhafte Werte der FDWH-Datensätze, die dennoch den Regeln (in Bezug zu Struktur und Datentypen) entsprechen, nur bedingt erkannt. Bei Fehlern und Inkonsistenzen in und zwischen Attributen mit Bezug zur semantischen Information des Datensatzes, gestaltet sich die Erkennung schwierig.

Die Ähnlichkeits- und Plausibilitätsanalyse hilft beim Erkennen von fehlerhaften bzw. nicht plausiblen Datensätzen oder Teilen davon, die durch Data Profiling nicht erkannt werden können. Dies geschieht durch den Vergleich mit anderen ähnlichen Lebensmittelprodukten mittels automatisierter Verfahren. Bei der Bewertung dieser Analysemethode wurden insgesamt 37 tatsächlich implausible Testdatensätze von insgesamt 100 Datensätzen aufgedeckt. Davon waren 26 Testdatensätze im Bereich der Inhaltsstoffe und 23 Testdatensätze im Bereich der Nährwerte implausibel. Das ROC-Diagramm zeigte außerdem, dass die Analyse mehrheitlich positiv klassifiziert wurde. Diese Art der Datenqualitätsanalyse wird somit als eine sinnvolle Methode zur Erkennung von fehlerhaften Datensätzen angesehen, auch wenn keine perfekte Klassifizierung möglich war. Oftmals war eine Bestimmung ähnlicher Produktdaten im FDWH nicht möglich, wenn das analysierte Produkt durch seine exotische Art nicht durch mehrere Datensätze in ähnlicher Form repräsentiert wurde. Dieser Fakt hat unter anderem negativ zur korrekten Klassifizierung während der ROC-Analyse beigetragen.

Die Ähnlichkeits- und Plausibilitätsanalyse ist sowohl im Bereich der Attribute für Inhaltsstoffe (der Inhaltsstoffliste) und Nährwerte, als auch für die Gesamtähnlichkeit möglich. Hier können zwar falsche Angaben erkannt, aber nicht in jedem Fall automatisch berichtigt werden. Möglich ist es, Berechnungen

einiger Nährwerteinformationen aus der Kombination anderer plausiblen Nährwerteinformationen durchzuführen. Dies ist beispielsweise bei den Nährwerteinformationen für Energie in Kilokalorien der Fall. Des Öfteren sind Korrekturen von implausiblen Datensätzen nur manuell durch Produktrecherchen möglich. Allerdings besteht die generelle Möglichkeit einer Übernahme von fehlenden Attributen eines Lebensmittelproduktdatensatzes von einem ihm ähnlichen Datensatz. Insbesondere im Bereich der Inhaltsstoffe können hierbei größere Abweichungen auftreten, wenn zwei Produkte zwar als ähnlich betrachtet werden, eines der Produkte in der Realität aber eine oder mehrere spezielle Zutaten beinhaltet, welche im anderen Produkt nicht enthalten sind. Das Produkt, welches die Inhaltsstoffe von ähnlichen Produkten übernimmt, übernimmt in diesem Fall mehr bzw. weniger Inhaltsstoffe, als es in der Realität beinhaltet. Aufgrund dieser Ungenauigkeiten während der Übernahme von Inhaltsstoffen muss vor der Verwendung diesbezüglicher Datensätze für risikobehaftete medizinische Einsatzgebiete gewarnt werden. Auch aufgrund fehlender Angaben von Inhaltsstoffen und Spuren von Zutaten, die zu weniger als 2 % im Produkt enthalten sind, kann keine Empfehlung der Lebensmittelproduktdaten für risikoreiche Anwendungszwecke ausgesprochen werden. Diese Sachverhalte gelten zum Beispiel für den Anwendungszweck im Bereich von Allergien und Nahrungsmittelunverträglichkeiten, wenn Anwender von Apps mit Zugriff auf das FDWH Risiken starker allergischer Reaktionen ausgesetzt sind (z. B. dem anaphylaktischen Schock) [59]. Eine Verantwortung liegt auch bei Betreuern und Therapeuten, die die jeweilige App mit Anbindung zum FDWH weiterempfehlen. Allerdings bietet die Ähnlichkeits- und Plausibilitätsanalyse die Möglichkeit vor fehlerbehafteten Daten zu warnen, indem implausible Datensätze erkannt werden. Weitere große Vorteile liefert die Ähnlichkeits- und Plausibilitätsanalyse sowie die Möglichkeit der Übernahme von Datenwerten in Anwendungsbereichen, bei denen eine Zufuhr von bestimmten Nährstoffen beobachtet werden muss und Abweichungen toleriert werden können. In diesem Zuge werden als Anwendungsbereiche die Symptome Untergewicht, Unterernährung, Essstörung, Adipositas und Fettstoffwechselstörungen (zum Teil auch Diabetes mellitus) als Beispiele genannt, bei denen entweder große Mengen oder möglichst wenige bestimmte Nährwerte verzehrt werden sollten.

Aber auch im Bereich der Lebensmittelallergien und -unverträglichkeiten können die genannten Methoden einen Mehrwert liefern. Hier gibt es beispielsweise bereits Ansätze, bei denen digitale Ernährungstagebücher durch statistische Methoden (Berechnung des relativen Risikos und Vierfelder-Tests, vgl. Rösch (2010) [59]) ausgewertet werden. In diesem Fall werden auch Inhaltsstoffe in die statistische Berechnung von Zusammenhängen zwischen verzehrten Lebensmitteln bzw. verzehrten Zutaten einbezogen, wenn ein Lebensmittel aus verschiedenen Inhaltsstoffen/Zutaten zusammengesetzt ist. Diese Methoden sollten allerdings nicht im Falle von Patienten mit kritischen allergischen Reaktionen eingesetzt werden [59]. Jede potenziell eingenommene Zutat eines Lebensmittels, insofern im jeweiligen Datensatz vorhanden, kann jedoch die Analyse des Zusammenhangs zwischen Lebensmittel und Symptom verbessern. Eine Kennzeichnung etwaiger Ungereimtheiten in den

Inhaltsstoffen (z. B. eine Kennzeichnung, dass Inhaltsstoffe von ähnlichem Datensatz übernommen wurden) kann in den Ergebnissen berücksichtigt werden. Z. B. kann dann eine Analyse der Daten mit den durch oben genannte Kennzeichnung versehenen Zutaten stattfinden und die gleiche Analyse durch Weglassen der gekennzeichneten Zutaten durchgeführt werden. Auffälligkeiten zwischen Inhaltsstoffen und Symptomen können anschließend in beiden Fällen verglichen und ausgewertet werden.

Zusammengefasst erleichtern die vom FDWH gelieferten Lebensmittel- und Lebensmittelproduktdaten den Ernährungsberatern und weiteren Verantwortlichen des Gesundheitswesens die Arbeit bei deren Analyse. Bei großen Mengen von Datensätzen ist der Aufwand der händischen Strukturierung und Überprüfung enorm groß, was durch die dargestellten automatisierten Verfahren stark vereinfacht wird. Des Weiteren wird die vereinfachte automatisierte Weiterverarbeitung durch (speziell dafür entwickelte) Gesundheitsapps vereinfacht, indem die Daten vereinheitlicht, einfach strukturiert und maschinenlesbar über eine API geliefert werden.

4.1.2 Der Einsatz von Big Data Technologie

Für die Speicherung sich zeitlich ständig verändernder Produktdaten, reichen relationale Datenbanken zur Datenspeicherung mit der Zeit nicht mehr aus. Des Weiteren ist eine händische Überprüfung der Qualität der Daten in solch großen Datenbeständen zeitlich nicht durchführbar. Im FDWH kommen diesbezüglich verschiedene Arten von Big Data Technologien zum Einsatz. TimescaleDB vermischt relationale Ansätze mit NoSQL-Ansätzen. Somit können durch Indizierung und Speicherung von Daten in mehreren Rechenclustern nicht nur große Mengen an zeitlich veränderlichen Produktdaten gespeichert werden. Auch das performante Selektieren dieser Daten ist möglich. Das Apache Spark Framework bringt Werkzeuge mit sich, um performant mehrere Datensätze miteinander zu vergleichen. Alle angewandten Big Data Lösungen skalieren Datensätze über mehrere Rechencluster hinweg. Obwohl diese Lösung durch die Art der Datenspeicherung sowie -verarbeitung bereits auf Single-Cluster-Systemen performanter arbeiten als herkömmliche Lösungen [B], erhöht sich die Performance mit jedem hinzugefügten Rechencluster, der mehr Rechenkapazität zur Verfügung stellt und parallel zu anderen Clustern im System arbeitet. Unter herkömmlichen Lösungen werden hierbei beispielsweise relationale Datenbanksysteme zur Datenspeicherung verstanden und das Verzichten von speziellen Big Data Frameworks wie Apache Spark. Dieses Framework verwendet sogenannte RDDs zur Zwischenspeicherung und Verarbeitung von Daten im Hauptspeicher und dem Mapping von Funktionen auf Datensätze. Durch die genannte Skalierung der Datenspeicherungs- und Datenverarbeitungssysteme können die beschriebenen Big Data Lösungen ihr Potenzial vollständig ausschöpfen. Dennoch fallen hierzu auch Mehrkosten in Bezug zur Beschaffung der Cluster-Hardware an, die bei der Architektur des Systems berücksichtigt werden müssen. Während der Test-Analysen des FDWHs im derzeitigen

Zustand, reichte die Anwendung der genannten Lösungen auf einem Singleclustersystem aus, um Daten über das FDWH ohne große Verzögerungen zur Verfügung zu stellen. Hierbei wurde als Konfiguration der Umgebung ein Testserver mit 8 x 3,5 GHz Rechenkernen und 64 GB Hauptspeicher zur Verfügung gestellt, der für die Anwendung des Apache Spark Frameworks ausreichend Ressourcen zur Verfügung stellte (siehe auch [B]). Durch Vergrößerung der im FDWH gespeicherten Datensätze beispielsweise durch Veränderungen an der Zusammensetzung von Produkten oder das Hinzufügen von Datensätzen und Datenquellen ist ggf. je nach Umfang der Daten eine Skalierung des Systems notwendig. Durch eigens durchgeführte Tests, bei denen eine in Python programmierte, „herkömmliche“ Datenanalyse-Anwendung durch das Apache Spark Framework ersetzt wurde, konnte eine Beschleunigung des Prozesses um das 90-fache beobachtet werden [B].

4.2 Anteil der einzelnen Prozessschritte im CRISP-Modell am Gesamtprozess

Wie in Chapman et al. (2000) beschrieben, nimmt bei Data Mining Projekten erfahrungsgemäß der Prozessteil der Datenaufbereitung mit ca. 50 % bis 70 % den größten Teil des Gesamtprojekt-Zeitaufwandes ein [15]. Des Weiteren werden nach Chapman et al. (2000) häufig 20 % bis 30 % des Zeitaufwandes für das Datenverständnis aufgebracht, 10 - 20 Prozent für die Modellierungsphase und 5 % bis 10 % für die Bereitstellungsphase [15]. Ähnlich zu diesen Anteilen des zeitlichen Aufwandes von dementsprechenden Projekten waren die Erfahrungen während der Entwicklungen der vorliegenden Arbeit. Während hierbei die Analysen für das Datenverständnis von den externen Datenquellen ca. 15 % der Entwicklungszeit einnahmen, verbrauchte die Phase der Datenvorbereitung etwa das Vierfache dieser Zeit. Die Regeln für die Transformation der Daten der unterschiedlichen Quellen mussten oft in mehreren Zyklen angepasst und dabei die jeweiligen Transformationsphasen wiederholt werden. Schrittweise wurde somit die Qualität der Daten für das Modelling anhand von Data Profiling und Transformationsschritten erhöht. Entwurf und Entwicklung während der Modellierungsphase nahmen ebenfalls ca. 15 Prozent des Zeitaufwandes in Anspruch. Der größte beanspruchte Teil bezog sich hierbei auf Recherche und das Experimentieren und Konfigurieren der verschiedenen Lösungsansätze. Die anschließende Phase der Evaluierung und Bereitstellung nahm die restlichen 10 % in Anspruch. Der größte Teil davon entfiel auf die manuelle Prüfung der Test-Lebensmittelproduktdatensätze anhand von Internetrecherchen und Prüfung von reellen Produktverpackungen.

4.3 Kommunikationsstandards und Standardisierung von Daten-Bezeichnern und -Ontologien der Lebensmittelproduktdaten

Wie häufig beim Umgang mit Datenobjekten mit Bezug zur Realität üblich, z. B. bei der Übertragung von Patientendaten in Krankenhäusern via Health Level 7 Standard (HL7), wäre eine Entwicklung eines Kommunikationsstandards bzw. der Standardisierung von Datenelementen und Ontologien der Lebensmittelproduktdaten für eine einfachere Verarbeitung in Gesundheitsanwendungen hilfreich. Eine Standardisierung von Bezeichnern der Angaben über Lebensmittelzusammensetzungen auf Produktverpackungen und in Produktdatenbanken würde eine Weiterverarbeitung der Daten unabhängig des jeweiligen Informationssystems vereinfachen [59]. Bezeichner von Zutaten könnten standardisiert werden, damit beispielsweise Zutaten wie die Orange (kann als Orange oder Apfelsine benannt werden) eine einheitliche Benennung in allen Datensätzen bekommen. Hierbei wäre auch eine Vereinheitlichung von Verarbeitungsschritten und der Angabe von Spuren in Lebensmitteln hilfreich. Dadurch würden auch Fehler in morphologischen Analysen durch Informationssysteme vermieden werden, da eine aufwendige Vorverarbeitung durch Text Mining Algorithmen entfällt [59].

4.4 Einsatz der beschriebenen Methoden für weitere mögliche Anwendungsfelder

Die eingesetzten Methoden der vorliegenden Arbeit wurden im Kontext von Lebensmitteldaten für medizinische Anwendungszwecke beschrieben. Datengrundlage für die Entwicklung von Analysemethoden sind Lebensmitteldaten bzw. Lebensmittelproduktdaten der FCDB und FPDB. Aufgrund der zur Verfügung stehenden Datengrundlage und notwendiger Anpassungen der Methoden an den jeweiligen Einsatzzweck sowie deren Einsatz im angewandten Forschungsbereich mit Bezug zur Ernährungsmedizin, wurde eine fachspezifische Ausrichtung der vorliegenden Arbeit gewählt. Dennoch ist der Einsatz der genannten Methoden, speziell der ETL-Prozess, die Data Profiling Methoden, die Big Data Technologie sowie die spezifisch entwickelten Ähnlichkeits- und Plausibilitätsanalyse-Methoden unter Einsatz von Data Mining Methoden für Projekte unterschiedlicher Gebiete und Absichten möglich. Eine Anpassung der Methoden an den jeweiligen Einsatzzweck ist notwendig. Die eigens entwickelte Ähnlichkeits- und Plausibilitätsanalyse kann zur Ermittlung der Plausibilität von verschiedenen Datenobjekten herangezogen werden. Fehler können hierbei in Datenobjekten ermittelt werden, die Eigenschaften beinhalten die sowohl durch Zeichenketten dargestellt als auch durch numerische Messwerte. Solche Zeichenketten enthalten beispielsweise standardisierte Bezeichnungen von vereinheitlichten Objektattributen, beispielsweise die Inhaltsstoffe in einer Inhaltsstoffliste von

Lebensmittelprodukten. Numerische Werte sind beispielsweise die Nährwerte der Lebensmittelprodukte. Zum Beispiel wäre der Einsatz der Ähnlichkeits- und Plausibilitätsanalyse auch bei anderen Arten von Produkten (anstelle von Lebensmittelprodukten) möglich. Die in Abschnitt 3.5.3 dargestellten Entscheidungsbäume der Plausibilitätsbestimmung müssten ggf. für den jeweiligen Anwendungszweck spezifisch generiert werden. Außerdem muss die Anzahl der durch das K-Means Verfahren erzeugten Clustern spezifisch angepasst werden. In diesem Zuge werden als Beispiel Smartphones als Technologie-Produkte in einem gänzlich zu Lebensmittelprodukten unterschiedlichem Anwendungskontext genannt. Angenommen, Datensätze mit Informationen zu Smartphones verschiedener Produktanbieter, die in einem zentralen Data-Warehouse gespeichert sind, sollen auf Plausibilität der Daten untersucht werden. Hierbei könnten, analog zu den Inhaltsstofflisten in Lebensmittelproduktdaten, Listen mit Produktspezifikationen (z.B. „matt schwarz, Qualcomm Snapdragon Prozessor, WiFi, 5G, Bluetooth, Fingerabdrucksensor“ [L45]) für eine Ähnlichkeitsanalyse von Zeichenketten verwendet werden. Als numerische Werte für die Ähnlichkeitsanalyse mittels Clustering können verschiedene Werte die technischen Messwerte der Bauteile des Smartphones beschreiben (z.B. 2,89 GHz, 8 GB RAM, 12 Megapixel, etc. [L45]). Anschließen ist es möglich, ähnliche Produkte durch die Ähnlichkeitsanalyse zu ermitteln und ggf. nach Anpassung der Entscheidungsbäume zur Plausibilitätsanalyse die Qualität der Daten zu untersuchen. Ein weiterer möglicher medizinischer Anwendungszweck, insbesondere zum Vergleich der Ähnlichkeit in Daten, wäre eine Analyse von Patientendaten. Hierbei könnten Beschreibungen zum Krankheitsverlauf, Medikationen etc. und verschiedene Messwerte von Biosignalen für die Analyse verwendet werden. Es wäre möglich, die Bestimmung der Plausibilität solcher Daten zu untersuchen. Diese Möglichkeit könnte zutreffen, wenn zwei Patienten gleiche Krankheitsbilder oder Messwerte, aber unterschiedliche Beschreibungen in weiteren Attributen wie beispielsweise der Medikation aufweisen würden. Bei dem Anwendungszweck müssten zwingend Datenschutzregelungen bezüglich personenbezogener Daten beachtet werden, um Rückschlüsse auf Einzelpersonen zu vermeiden, was bei der Analyse von Produktdaten nicht notwendig ist. Tabelle 4.1 zeigt eine Übersicht der soeben erläuterten sowie weiteren Beispiele für eine Anwendung der Ähnlichkeits- und Plausibilitätsanalyse.

Art der Datensätze	**Beispiel**
Technische Produktdaten	Smartphones, Laptops, Pkw
Medizinische Patientendaten	Anamnese- und Diagnosedaten von Patienten
Medizinische Produktdaten	Medikamente, technische Medizinprodukte
Ökonomische Daten	Marktforschungsdaten
Landwirtschaftliche Daten	Landwirtschaftliche Prognosen, Daten über Wachstumsrate, Bodeneigenschaften etc. von Feldern
Naturwissenschaftliche Daten	Daten über biologische oder chemische Vorgänge

Tabelle 4.1 Beispiele von Daten für eine Anwendung der Ähnlichkeits- und Plausibilitätsanalyse

Insgesamt lässt sich sagen, dass eine ähnliche Form der in dieser Arbeit beschriebenen Ähnlichkeits- und Plausibilitätsanalyse dann möglich ist, wenn Datensammlungen Datensätze mit numerischen Werten und vereinheitlichte Bezeichner von Eigenschaften der Datenobjekte als Zeichenkette beinhalten. Wie beschrieben, müssen die Data Mining Methoden zur Klassifizierung angepasst werden. Dies betrifft die Anzahl an Clustern, die durch das K-Means Verfahren erzeugt werden sowie die Entscheidungsbäume für die Plausibilitätsklassifizierung.

5 Zusammenfassung

In der vorliegenden Arbeit wird ein zentrales Lebensmittel-Datawarehouse, auch FDWH genannt, entwickelt. Der Vorgang dabei gestaltet sich nach den im CRISP-DM-Modell beschriebenen Prozesszyklen. Diese beziehen sich auf die einzelnen Prozessschritte, die während Data Mining Projekten durchgeführt werden. Zunächst erfolgt eine Analyse der potenziellen Anwendungszwecke und der Datenbasis in Form verschiedener externer Datenquellen, die teilweise durch Internet-Communitys gepflegt werden, sowie deren Meta- und Wissensdaten. Anschließend werden die Daten für die Speicherung im FDWH vorbereitet. Mithilfe von Ontologien und Data Profiling werden Verfahren und Regeln für die Transformation während des ETL-Prozesses für die konsistente und vereinheitlichte Speicherung erarbeitet und implementiert.

Durch den ETL-Prozess konnten Daten aus externen Datenquellen extrahiert werden, die sich in verschiedenen Datenformaten befinden. Anschließend werden die Daten in das Format bzw. die Datenstruktur des FDWH transformiert und in dessen zentrale Datenbank geladen. Die Data Profiling Regeln ermöglichen es Datenstruktur und Datentypen von Daten aus externen Quellen automatisiert zu transformieren. Durch das Bilden von eindeutigen Hashwerten als IDs für die Datentabellen mit Informationen zu den Lebensmitteln, werden doppelte Datensätze leicht erkannt. Lebensmittelproduktdatensätze, die die gleiche ID enthalten, sind somit doppelt im FDWH enthalten. Die Ähnlichkeits- und Plausibilitätsanalyse wird eingesetzt, um Daten auf Plausibilität zu untersuchen und fehlende Werte zu vervollständigen. Data Mining Verfahren, wie K-Means, Entscheidungsbäume und die Kosinus-Distanz-Berechnung werden angewendet, um Ähnlichkeiten in Datensätzen zu erkennen, Datensätze zu vergleichen und je nach Plausibilität und Implausibilität zu klassifizieren. Fehlende Daten in den Datensätzen können aus ermittelten, ähnlichen Datensätzen übernommen werden. Durch den Einsatz der Big Data Technologie TimescaleDB können große Datenmengen performant gespeichert werden, die beim Einfügen von sich über die Zeit veränderten Datensätze ins FDWH anfallen. Das PySpark Framework erlaubt eine performante Analyse durch Vergleichen größerer Mengen von Datensätzen und Datenattributen. Neben den genannten Analysemethoden wurde für Tests eine grafische Weboberfläche implementiert, anhand derer automatisiert auf Qualität geprüfte Produktdatensätze im FDWH über eine Produktsuche selektiert und grafisch aufbereitet für eine anschließende manuelle Überprüfung angezeigt werden können. Über diese Weboberfläche können zusätzlich Details über den Anwendungskontext mitgegeben werden. Details sind in dem Fall Inhaltsstoffe, Nährwerte oder Allergene deren Vorhandensein in einem Produktdatensatz nach der Abfrage überprüft und das Ergebnis angezeigt wird. Die Lebensmitteldaten des FDWH werden durch Mapping- und Cluster-Verfahren kategorisiert.

Die Bewertung der Data Profiling Methoden hat gezeigt, dass die Methoden des Data Profilings Inkonsistenzen durch das Anwenden bestimmter Regeln ermitteln und korrigieren können. Dies trifft

aber nur in dem Fall zu, wenn eine Regel existiert, die für eine Erkennung des jeweiligen Sachverhalts der Inkonsistenz geeignet ist. Somit können Inkonsistenzen auch nur dann bereinigt werden, wenn bestimmte Regeln zur Umwandlung oder der Berechnung eines Wertes anhand anderer Werte (desselben Datensatzes) festgelegt wurden. Die Ähnlichkeits- und Plausibilitätsanalyse ist dazu geeignet, Inkonsistenzen durch Vergleichen zueinander ähnlicher Datensätze zu erkennen und fehlende Werte zu ermitteln. Die gültige Funktion dieser Art der Analyse wurde durch die Evaluierung in Abschnitt 3.8.2 bestätigt. Die entwickelte Plausibilitäts- und Ähnlichkeitsanalyse erweitert die zur Verfügung stehenden Möglichkeiten der Qualitätsverbesserung durch Data Profiling und liefert somit einen Mehrwert im Bereich der Qualitätsoptimierung. Die Analysemethode kann auch auf andere Anwendungskontexte außerhalb von Lebensmittelproduktdaten angewandt werden. Dies erfordert jedoch eine Anpassung der Algorithmen. Zu beachten im Umgang mit Lebensmittelproduktdaten gilt, dass falsche Angaben in den Daten schwerwiegende Folgen für Konsumenten haben können. Daten, die nicht zweifelsfrei ohne Inkonsistenzen und Fehler sind, können deshalb nicht mit Gewähr an Gesundheits-apps geliefert werden. Dennoch konnte durch die vorliegende Arbeit eine Datenbasis mit Lebensmittel- und Lebensmittelproduktdaten aufgebaut werden, deren Anwendung für viele ernährungsbasierte Forschungszwecke möglich ist.

Literaturverzeichnis

[1] Abedjan, Z., Golab, L., & Naumann, F. (2016). Data Profiling. 2016 IEEE 32nd International Conference on Data Engineering (ICDE), S. 1432-1435. doi: 10.1109/ICDE.2016.7498363.

[2] Altman, D., & Bland, J. (1994). Diagnostic tests. 1: Sensitivity and specifity. BMJ: British Medical Journal., 308, Nr. 6943, S. 1552.

[3] Apel, D., Behme, W., Eberlein, R., & Merighi, C. (2015). Datenqualität erfolgreich steuern: Praxislösungen für Business-Intelligence-Projekte. 3. Auflage, Abs. 9.4. Heidelberg: dpunkt.verlag.

[4] Arens, A., Rösch, N., Feidert, F., Harpes, P., Herbst, R., & Mösges, R. (2008). Mobile electronic patient diaries with barcode based food identification for the treatment of food allergies. GMS Medizinische Informatik, Biometrie und Epidemiologie, 4.

[5] Bankhofer, U. V. (2008). Entscheidungsbäume. In *Datenanalyse und Statistik*, S. 273-284. Wiesbaden: Gabler. doi:10.1007/978-3-8349-9654-1_18

[6] Bansal, S. K., & Kagemann, S. (2015). Integrating Big Data: A Semantic Extract-Transform-Load Framework. Computing in Science & Engineering, vol. 48, no. 3, S. 42-50. doi:10.1109/MC.2015.76.

[7] Barth, S., & Kraft, M. (2009). Ernährungsmedizin BASICS. München: Elsevier GmbH, Urban & Fisher Verlag.

[8] Bartolomei, T. T., Czarnecki, K., Lämmel, R., & van der Storm, T. (2009). Study of an API migration for two XML APIs. In van den Brand M., Gašević D., Gray J. (eds) Software Language Engineering. SLE 2009, Lecture Notes in Computer Science, vol 5969. Berlin, Heidelberg: Springer. doi: 10.1007/978-3-642-12107-4_5

[9] Bauer, A., & Günzel, H. (2013). Metadaten. In Data Warehouse Systeme, S. 339-368. Heidelberg: dpunkt.verlag.

[10] Bauer, A., & Günzel, H. (2013). Referenzarchitektur. In: Data Warehouse Systeme. S. 37-86. Heidelberg: dpunkt.verlag.

[11] Bauer, A., & Günzel, H. (2013). Umsetzung des multidimensionalen Datenmodells. In: Data Warehouse Systeme S. 241-297. Heidelberg: dpunkt.verlag.

[12] Biesalski, H. K., Grimm, P., & Nowitzki-Grimm, S. (2020). Ernährungsmedizin. In *Taschenatlas Ernährung*, 8 Ausg., S. 374-413. Stuttgart: Thieme Verlag.

[13] Bradley, A. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithm. *Pattern Recognition*, Volume 30, S. 1145-1159.

[14] Cánovas Izquierdo, J. L., & Cabot, J. (2013). Discovering Implicit Schemas in JSON Data. In *In: Daniel F., Dolog P., Li Q. (eds) Web Engineering. ICWE 2013. Lecture Notes in Computer Science* Vol. 7977, S. 68-83. Berlin, Heidelberg: Springer. doi: 10.1007/978-3-642-39200-9_8

[15] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0.* USA: SPSS Inc.

[16] Chauhan, D., & Bansal, K. L. (2017). Using the Advantages of NOSQL: A Case Study on MongoDB. *International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC)*, S. 90-93.

[17] Cleve, J., & Lämmel, U. (2020). Klassifikation. *In: Data Mining*, 3. Auflage, S. 87-147. Oldenburg: De Gruyter.

[18] Cleve, J., & Lämmel, U. (2020). Ablauf einer Datenanalyse. *In: Data Mining* 3. Auflage, S. 2-7. Oldenburg: De Gruyter.

[19] Cleve, J., & Lämmel, U. (2020). Interdisziplinarität von Data Mining. In *In: Data Mining* (3. Auflage Ausg., S. 11-15). Oldenburg: De Gruyter.

[20] Delone, W. H., & McLean, E. R. (1992). Information System Success: The Quest for the Dependent Variable. *In: Information System Research, 3. Jg., Nr. 1 (1992), S. 60–95.*

[21] Dengel, A., & Bernardi, A. (2012). Metadaten. In A. Dengel, *Semantische Technologien*, S. 13-15. Heidelberg: Springer, Spektrum Akademischer Verlag.

[22] Dengel, A., & Bernardi, A. (2012). Semantik und semantische Technologien. In A. Dengel, *Semantische Technologien*, S. 10-13. Heidelberg: Springer, Spektrum Akademischer Verlag.

[23] Dengel, A., Bernardi, A., & van Elst, L. (2012). Wissensrepräsentation. In A. Dengel, *Semantische Technologien*, S. 23-72. Heidelberg: Springer, Spektrum Akademischer Verlag.

[24] Dig, D., & Johnson, R. (2006). How do APIs evolve? A story of refactoring. *Journal of Software Maintenance and Evolution Research and Practice. John Wiley & Sons, Ltd.*

[25] Dunteman, G. H. (1989). *Principal Component Analysis.* Kalifornien, USA: Sage Publications.

[26] Elfert, P., et al. (2017). DiDiER - digitized services in dietary counselling for people with increased health risks related to malnutrition and food allergies. *2017 IEEE Symposium on Computers and Communications (ISCC)*, S. 100-104. doi:10.1109/ISCC.2017.8024512

[27] Evers-Wölk, M., Oertel, B., & Sonk, M. (2018). *GesundheitsApps, Innovationsanalyse.* Büro für Technikfolgen-Abschätzung (TAB) beim deutschen Bundestag, Berlin. Bad Honnef: Wienans Print + Medien GmbH.

[28] Eysenbach, G. (2001). What is E-Health? *Editorial of the Journal of Medical Internet Research*, S. 3(2):e20.

[29] Fawcett, T. (2003). *ROC Graphs: Notes and Practical Considerationsfor Data Mining Researchers.* Intelligent Enterprise Technologies Laboratory, HP Laboratories Palo Alto, HPL-2003-4. [Online] Zuletzt abgerufen am 30. Oktober 2021. URL: https://www.hpl.hp.com/techreports/2003/HPL-2003-4.pdf

[30] Garani, G., & Helmer, S. (2012). Integrating Star and Snowflake Schemas in Data Warehouses. *22 International Journal of Data Warehousing and Mining*, S. 8(4), 22-40.

[31] Gigerenzer, G., Schlegel-Matthies, K., & Wagner, G. G. (2016). *Digitale Welt und Gesundheit. eHealth und mHealth – Chancen und Risiken der Digitalisierung im Gesundheitsbereich* . SVRV Sachverständigenrat für Verbraucherfragen. [Online] Zuletzt abgerufen am 30. Oktober 2021. URL: https://www.svr-verbraucherfragen.de/wp-content/uploads/Digitale-Welt-und-Gesundheit.pdf

[32] Glushko, R. J., Tenenbaum, J. M., & Meltzer, B. (1999). AN XML FRAMEWORK FOR Agent-based E-commerce. *COMMUNICATIONS OF THE ACM*.

[33] Grzesik , P., & Mrozek , D. (2020). Comparative Analysis of Time Series Databases in the Context of Edge Computing for Low Power Sensor Networks. *In: Krzhizhanovskaya V.V. et al. (eds) Computational Science – ICCS 202,. Lecture Notes in Computer Science, vol 12141*. doi: 10.1007/978-3-030-50426-7_28

[34] GS1 Germany GmbH. (2009). *Die GS1 Nummernsysteme.* GS1Tech. [Online] Zuletzt abgerufen am 21. September 2020 URL: https://www.gs1-germany.de/fileadmin/gs1/basis_informationen/nummernsysteme_grundlage_weltw_daten_warenverkehrs.pdf

[35] Helfert, M. (2002). *Planung und Messung der Datenqualität in Data-Warehouse Systemen.* Bamberg: Difo-Druck GmbH.

[36] Imasheva, B., Nakispekov, A., Sidelkovskaya, A., & Sidelkovskiy, A. (2020). The Practice of Moving to Big Data on the Case of the NoSQL Database, Clickhouse. In: *Optimization of Complex Systems: Theory, Models, Algorithms and Applications,* S. 820-828. Springer International Publishing. doi:10.1007/978-3-030-21803-4_82

[37] Kälin, P. Hauptkomponentenanalyse PCA. ETH Zürich. [Online] Zuletzt abgerufen am 24. Oktober 2021. *URL:* https://www.analytik.ethz.ch/vorlesungen/chemometrie/2_PCA_Printer.pdf

[38] Keuthage, W., & Schoppe, T. (2016). Ernährung goes digital. In: (B. Kluthe, E. Linker, R. Obeid, E. Purucker, & K. Schmidt, Hrsg.) *E&M - Ernährung und Medizin*, S. 124-128.

[39] Kimball, R., Reeves, L., Ross, M., & Thornwaite, W. (1998). Collecting the Requirements. In: The Data Warehouse Lifecycle Toolkit, *Abschnitt 3.3.* New York (USA): John Wiley & Sons.

[40] Kraft, T. (2017). *Masterarbeit - Qualitätsmaße binärer Klassifikatoren im Bereich kriminalprognostischer Instrumente der vierten Generation.* Algorithm Accountability Lab, Fachbereich Informatik, Technische Universität Kaiserslautern. arXiv:1804.01557v1 [cs.CY]

[41] Krause, T. (2012). *Diplomarbeit - Entwurf und Implementierung einer effizienten Dublettenerkennung für großer Adressbestände.* Fachhochschule Köln, Fakultät für Informatik und Ingenieurswissenschaften.

[42] Kröger, G. *eHealth und Big Data im Gesundheitswesen: Analyse zur PWC-Studie und des Charismha-Projekts; Hrsg. Brendan-Schmittmann-Stiftung.*

[43] Kusumasari, T., & Fitria. (2016). Data Profiling for Data Quality Improvement with Openrefine. *IEEE International Conference on Information Technology Systems and Innovation (ICITS)*, S. 1-6. doi: 10.1109/ICITSI.2016.7858197

[44] Laney, D. (2001). *3D Data Management: Controlling Data Volume, Velocity and Variety*. META Group Research Note, 6. Gartner.

[45] Li, Y., Yu, X., & Li, C. (2010). The Applied Research on the Statistic Data Warehouse Based on the Snowflake Mode. *2nd International Workshop on Database Technologie and Applications*, S. 1-4. doi:10.1109/DBTA.2010.5659083

[46] Los, R. (2006). *Supporting Uniform Representation of Data: Structuring Medical Narratives for Care and Research.* Erasmus University Rotterdam. [Online] Zuletzt abgerufen am 27. Juli 2021 URL: http://hdl.handle.net/1765/7579

[47] Martin, E. (2007). *100 Fragen zur Ernährung kranker Menschen.* Hannover: Schlütersche Verlagsgesellschaft mbH & Co. KG.

[48] Matthiessen, G., & Unterstein, M. (2008). Datenbankentwurf. In: *Relationale Datenbanken und Standard-SQL, 4. Auflage*, S. 91-160. München: Addison-Wesley.

[49] Mitlöhner, J., Neumaier, S., Umbrich , J., & Polleres, A. (2016). Characteristics of Open Data CSV Files. *2016 2nd International Conference on Open and Big Data (OBD)*, S. 72-79. doi:10.1109/OBD.2016.18

[50] Müller, D. (2015). *Masterarbeit; Apache Spark: Untersuchung der Möglichkeiten zur verteilten Datenverarbeitung und Analyse von Streaming Data.* Offenburg: Studiengang Informatik (INFM) an der Fakultät Elektrotechnik und Informationstechnik der Hochschule für angewandte Wissenschaften Offenburg.

[51] Ng, A., & Soo, K. (2018). Hauptkomponentenanalyse. In: *Data Science – was ist das eigentlich?!* S. 29-43. Berlin, Heidelberg: Springer. doi: 10.1007/978-3-662-56776-0_3

[52] Olson, J. (2003). *Data Quality: The Accuracy Dimension.* Morgan Kaufmann Publishers.

[53] Oussous, A., Benjelloun, F.-Z., Lahcen, A. A., & Belfkih, S. (2017). NoSQL databases for big data. *Big Data Intelligence, Volume 4, Issue 3*, S. 171-185.

[54] Patil, D. D., Wadhai, V. M., & Gokhale, J. A. (2010). Evaluation of Decision Tree Pruning Algorithms for Complexity and Classification Accuracy. *International Journal of Computer Applications, Volume 11– No.2*, S. 23-30.

[55] Pinto, D., Vilarino, Y., Aleman, Y., Gomez, H., & Loya, N. (2012). The Soundex Phonetic Algorithm Revisited for SMS-based Information Retrieval. *Conferencia Espanola de la Recuperacion de Informacion (CERI 2012)*.

[56] Pirrung, M. (2020). *Analyse zum Einsatz von Gesundheits-Apps beim Management von Nahrungsmittelunverträglichkeit und -allergie*. Zweibrücken: Masterarbeit, Applied Life Sciences, Hochschule Kaiserslautern.

[57] Prokosch, H. U. (2001). KAS, KIS, EKA, EPA, EGA, E-Health: Ein Plädoyer gegen die babylonische Begriffsverwirrung in der Medizinischen Informatik. Informatik, Biometrie und Epidemiologie in Medizin und Biologie 32,4, S. 371-382.

[58] Rehs, P. H. (2014). *Bachelorarbeit: Verfahren zur Dimensionsreduktion.* Düsseldorf: Institut für Informatik, Datenbanken und Informationssysteme, Heinrich Heine Universität.

[59] Rösch, N. (2010). *Der Einsatz von Informations- und Kommunikationstechnologie bei Nahrungsmittelallergie.* Aachen: Shaker Verlag.

[60] Rösch, N., Arens-Volland, A., Harpes, P., Herbst, R., Plumer, P., Feidert, F., . . . Mösges, R. (2009). Telemedizinisch unterstütztes Diät- und Diagnosemanagement bei Nahrungsmittelallergie. *BULLETIN de la Société des Sciences Médicales du Grand-Duché de Luxembourg*, S. 163-170.

[61] Rösch, N., Feidert, F., Arens, A., & Mösges, R. (2007). MENSSANA Mobile Expert & Networking System for Systematical Analysis of Nutrition based Allergies. *Allergy (European Journal of Allergy and Clinical Immunology), 62*, S. 565-566.

[62] Rösch, N., Münzberg, A., Sauer, J., Arens-Volland, A., Lämmel, S., Teichmann, S., Eichelberg, M., Hein, A. (2019). Digital supported diagnostics in food allergy by analyzing app-based diaries. *Allergy Volume 74 Issue S1006 Special Issue: Abstracts from the European Academy of Allergy & Clinical Immunology (EAACI) Congress*. doi: 10.1111/all.13959

[63] Rutz, M., Kühn, D., & Dierks, M. L. (2016). *Gesundheits-Apps und Prävention.* In: Chancen und Risiken von Gesundheits-Apps (CHARISMHA). Braunschweig: Peter L. Reichertz Institut für Medizinische Informatik der TU Braunschweig und der Medizinischen Hochschule Hannover. [Online] Zuletzt abgerufen am 21. Juli 2020 URL: http://www.digibib.tu-bs.de/?docid=00060010

[64] Sauer, J., Hein, A., Muenzberg, A., & Roesch, N. (2019). Simplify Testing of Mobile Medical Applications by Using Timestamps for Remote, Automated Evaluation. *2019 International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, S. 203-206. doi:10.1109/WiMOB.2019.8923241

[65] Sauer, J., Muenzberg, A., Siewert, L., Hein, A., & Roesch, N. (2019). Remote Testing of Usability in Medical Apps. *Wireless Mobile Communication and Healthcare. 8th EAI International Conference, MobiHealth 2019, Dublin, Ireland, November 14-15, 2019, Proceedings*. doi: 10.1007/978-3-030-49289-2_1

[66] Sawade, C., Landwehr, N., Prasse, P., Makowski, S., & Scheffer, T. *Entscheidungsbäume.* Universität Potsdam, Institut für Informatik, Lehrstuhl

Maschinelles Lernen. [Online] Zuletzt abgerufen am 24. Oktober 2021 URL: https://www.cs.uni-potsdam.de/ml/teaching/ws11/ml/Entscheidungsbaeume.pdf

[67] Simons, M. (2018). Spring Boot Grundlagen. In M. Simons, *In: Spring Boot 2: Moderne Softwareentwicklung mit Spring 5,* S. 53-64. Heidelberg: dpunkt.verlag GmbH.

[68] Simons, M. (2018). Spring Web-MVC. In M. Simons, *In: Spring Boot 2: Moderne Softwareentwicklung mit Spring 5,* S. 53-64. Heidelberg: dpunkt.verlag GmbH.

[69] Stuckenschmidt, H. (2011). Ontologien Konzepte, Technologien und Anwendungen. In O. P. Günther, W. Karl, R. Lienhart, & K. Zeppenfeld (Hrsg.), *Informatik im Fokus,* 2. Auflage, S. 25-51. Heidelberg: Springer.

[70] Tantau, T. (2014). *Kapitel 1.* Syntax versus Semantik: Text und seine Bedeutung. Universität zu Lübeck. [Online] Zuletzt abgerufen am 24. Oktober 2021 URL: http://www.tcs.uni-luebeck.de/downloads/mitarbeiter/tantau/svgtest/presentation_version.pdf

[71] Tilkov, S., Eigenbrodt, M., Schreier, S., & Wolf, O. (2015). Einleitung. In *In: REST und HTTP - Entwicklung und Integration nach dem Architekturstil des Web,* 3. Auflage, S. 1-5). Heidelberg: dPunk.Verlag.

[72] Vukotic, A., & Goodwill, J. (2011). Chapter 1: Introducing to Apache Tomcat 7. In A. Vukotic, & J. Goodwill, *Apache Tomcat 7,* 1. Ausg., S. 1-6. Apress.

[73] Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *In: Journal of ManagementInformation Systems, 12. Jg., Nr. 4 (1996), S. 5–33.*

[74] Ward, J. S., & Barker, A. (2013). *Undefined By Data: A Survey of Big Data Definitions.* arXiv: abs/1309.5821

[75] Wissuwa, S., Cleve, J., & Lämmel, U. (2005). Analyse zeitabhängiger Daten durch Data-Mining-Verfahren. *In: Wismarer Diskussionspapiere, Hochschule Wismar, Fachbereich Wirtschaft*.

[76] Wolff, E. (2010). Aspektorientierte Programmierung mit Spring. In E. Wolff, *In: Spring 3, Framework für die Java Entwicklung,* 3. Ausg., S. 99-142. Heidelberg: dpunkt.verlag GmbH.

[77] Wolff, E. (2010). Dependency Injection. In E. Wolff, *In: Spring 3, Framework für die Java Entwicklung,* 3. Ausg., S. 11-98. Heidelberg: dpunkt.verlag GmbH.

[78] Wolff, E. (2010). Spring Webtechnologien. In E. Wolff, *In: Spring 3, Framework für die Java Entwicklung*, 3. Ausg., S. 231-298. Heidelberg: dpunkt.verlag GmbH.

[79] Worm, M. (2018). Anaphylaxie: Wie richtig handeln? *Deutsches Ärzteblatt 2018*, S. 115(10). doi:10.3238/PersPneumo.2018.03.09.02

[80] Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. (2010). Spark: Cluster Computing with Working Sets. *HotCloud 2010*.

Vorabveröffentlichungen

[A] Münzberg, A., Sauer, J., Hein, A., & Rösch, N. (2018). The use of ETL and data profiling to integrate data and improve quality in food databases. *14th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, S. 231-238. doi:10.1109/WiMOB.2018.8589081

[B] Münzberg, A., Sauer, J., Hein, A., & Rösch, N. (2019). Checking the Plausibility of Nutrient Data in Food Datasets Using KNIME and Big Data. *International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, S. 1-4. doi:10.1109/WiMOB.2019.8923233

[C] Münzberg, A., Sauer, J., Lämmel, S., Teichmann, S., Hein, A., & Rösch, N. (2019). Optimization and merging of food product data and food composition databases for medical use. *European Academy of Allergy & Clinical Immunology (EAACI) Congress*.

[D] Münzberg, A., Sauer, J., Hein, A., & Rösch, N. (2020). Entwicklung eines Data Warehouse mit Lebensmittelproduktdaten für Gesundheits-Apps. In M. Sprenger, C. Dindorf, S. Defren, B. Steinke, & M. Fröhlich (Hrsg.), *Sports, Movement & Health*, Bd. 1, S. 99-100. Kaiserslautern: Technische Universität Kaiserslautern.

[E] Münzberg, A., Sauer, J., Hein, A., & Rösch, N. (2020). Intelligent Combination of Food Composition Databases and Food Product Databases for Use in Health Applications. (G. O'Hare, M. O'Grady, J. O'Donoghue, & P. Henn, Hrsg.) *Wireless Mobile Communication and Healthcare. 8th EAI International Conference (MobiHealth)*. doi: 10.1007/978-3-030-49289-2

[F] Muenzberg, A., Sauer, J., Hein, A., & Roesch, N. (2021). Machine Learning and Context-based Approaches to Get Quality Improved Food Data. *Yang, X.-S., Sherrat, S., Dey, N., Joshi, A. (Eds.). Sixth International Congress on Information and Communication Technology (ICICT 2021). London, Volume 2, S. 423-435.*

Internetverweise

[L1] 42matters AG. (2020). *Google Play vs the iOS App Store: Store Stats for Mobile Apps*. Zuletzt abgerufen am 18. Juli 2020. URL: https://42matters.com/stats

[L2] Alnatura Produktions- und Handels GmbH. (2021). *Bio-Siegel und -Verbände*. Zuletzt abgerufen am 27. Juli 2021. URL: https://www.alnatura.de/de-de/ueber-uns/bio-siegel-und-verbaende/

[L3] AnyConnector. (2021). *AnyConnector*: Was ist Data Profiling? Definition, Techniken und Vorteile. Zuletzt abgerufen am 24. Oktober 2021 URL: https://anyconnector.com/de/data-transformation/what-is-data-profiling.html

[L4] Ballard, C. (2019). Medium: *How We Use Machine Learning to Turn Product Packaging into Structured Data*. Zuletzt abgerufen am 2. September 2021 URL: https://medium.com/nielsen-forward/how-we-use-machine-learning-to-turn-product-packaging-into-structured-data-ba223c3bff20

[L5] Brucker, B. (2019). *Trend zur Selbstvermessung: Chancen und Risiken von Gesundheits-Apps.* Zuletzt abgerufen am 11. Juni 2020 URL: https://www.sifa-sibe.de/gesundheit/gesundheits-apps/

[L6] Bundesministerium für Ernährung und Landwirtschaft (BMEL). (2019). *Allergenkennzeichnung ist Pflicht.* Zuletzt abgerufen am 11. Juli 2021 URL: https://www.bmel.de/DE/themen/ernaehrung/lebensmittel-kennzeichnung/pflichtangaben/allergenkennzeichnung.html

[L7] Bundesnetzagentur. (2016). *Bundesnetzagentur für Elektrizität, Gas, Telekommunikation, Post und Eisenbahnen:* Bekanntmachung zur elektronischen Signatur nach dem Signaturgesetz und der Signaturverordnung. Zuletzt abgerufen am 23. Mai 2021. URL: https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/ElekSignatur/Algorithmenkatalog2017_Entwurf.pdf?__blob=publicationFile&v=4

[L8] Beuth Verlag GmbH. (2020). *Beuth publishing DIN*: DIN EN ISO 14971:2020-07. Zuletzt abgerufen am 24. Oktober 2021 URL: https://www.beuth.de/de/norm/din-en-iso-14971/307629814

[L9] Castillo, G., & Jorzyk, K. *Definition Information:* Identität und Geschichte der Informationswissenschaft, Informationswissenschaftliche Themen. Zuletzt abgerufen am

09. Mai 2021 URL: https://saar.infowiss.net/projekte/ident/themen/definition-information/

[L10] Chamoni, P. (2012). *Data Mining*. Zuletzt abgerufen am 21. September 2020 URL: https://www.enzyklopaedie-der-wirtschaftsinformatik.de/wi-enzyklopaedie/lexikon/daten-wissen/Business-Intelligence/Analytische-Informationssysteme--Methoden-der-/Data-Mining/index.html

[L11] CodeCheck AG. *Shop better live better*. Zuletzt abgerufen am 28. August 2021 URL: codecheck: https://codecheck-app.com/

[L12] DANONE GmbH. *DANONE*. Zuletzt abgerufen am 11. September 2021 URL: https://www.danone.de/geschaeftsbereiche/danone-essential-dairy-plant-based-products/unsere-werke.html

[L13] Deutsches Institut für Normung e. V. (DIN). (2015). *DIN-Normenausschuss Qualitätsmanagement, Statistik und Zertifizierungsgrundlagen (NQSZ)*: ISO 9001 Qualitätsmanagementsysteme - Anforderungen. Zuletzt abgerufen am 11. September 2021. URL: https://www.din.de/de/mitwirken/normenausschuesse/nqsz/veroeffentlichungen/wdc-beuth:din21:242367583

[L14] Europäische Union (EU). (2017). *VERORDNUNG (EU) 2017/745 DES EUROPÄISCHEN PARLAMENTS UND DES RATES*. Zuletzt abgerufen am 27. Juli 2021 URL: https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX%3A32017R0745

[L15] Fink, L. (2019). *Hidden treasures in our groceries*. Zuletzt abgerufen am 06. Juni 2021 URL: kaggle.com: https://www.kaggle.com/allunia/hidden-treasures-in-our-groceries

[L16] Göbel, A. (2017). *regexe*. Zuletzt abgerufen am 4. September 2020 URL: http://www.regexe.de/hilfe.jsp

[L17] Graphite. Graphite. Zuletzt abgerufen am 22. Oktober 2021 URL: http://graphiteapp.org/

[L18] Grootendorst, M. (2021). *9 Distance Measures in Data Science*. Zuletzt abgerufen am 15. August 2021 URL: https://www.maartengrootendorst.com/blog/distances/

[L19] Grünwald, R. (2019). *Data Mining Klassifikation: Entscheidungsbaum und neuronale Netze gewinnbringend nutzen!* Zuletzt abgerufen am 15. August 2021 URL: https://novustat.com/statistik-blog/data-mining-klassifikation-gewinnbringend.html

[L20] GS1 Germany GmbH. *GS1 Germany*: Global Trade Item Number. Zuletzt abgerufen am 11. September 2021 URL: https://www.gs1-germany.de/gs1-

standards/identifikation/artikel-gtin-sgtin/#:~:text=Aufbau%20der%20GTIN,die%20Artikelnummern%20f%C3%BCr%20Ihre%20Produkte.

[L21] Gupta, P. (2017). *Decision Trees in Machine Learning*. Zuletzt abgerufen am 18. Juni 2021 URL: https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052

[L22] Imru, I. (2020). *Medium*: Time Series Database Comparison. Zuletzt abgerufen am 22. Oktober 2021 URL: https://medium.com/griddb/time-series-database-comparison-aa83c5e257d2

[L23] International Organization for Standardization (ISO). (2002). *ISO 639-1:2002*. Zuletzt abgerufen am 11. September 2021 URL: https://www.iso.org/standard/22109.html

[L24] International Organization for Standardization (ISO). (2009). *ISO 80000-1:2009*. Zuletzt abgerufen am 11. September 2021 URL: https://www.iso.org/standard/30669.html

[L25] International Organization for Standardization (ISO). (2013). *ISO 3166-1:2013*. Zuletzt abgerufen am 11. September 2021 URL: https://www.iso.org/standard/63545.html

[L26] Johner, C. (2021). *Johner Institut*: Software-Risikomanagement für medizinische Software. Zuletzt abgerufen am 2024. Oktober 2021 URL: https://www.johner-institut.de/blog/iec-62304-medizinische-software/software-risikomanagement-iso14971/

[L27] Klemm, S. (2018). Medium: *Grafana & TimescaleDB: enhancing time-series exploration and visualization*. Zuletzt abgerufen am 13. September 2021 URL: https://medium.com/timescale/grafana-time-series-exploration-visualization-postgresql-8c7baa9c3bfe

[L28] Klöckner, L. (2021). *Onmeda.de*. (F. D. GmbH, Herausgeber): Kann man Kalorienangaben auf Lebensmitteln trauen? Zuletzt abgerufen am 20. Oktober 2021 URL: https://www.onmeda.de/ernaehrung/kalorienangaben.html#aussagekraft

[L29] KNIME AG. *End to End Data Science*. Zuletzt abgerufen am 13. September 2021 URL: https://www.knime.com/

[L30] Konradin Medien GmbH. (2018). *Mikronährstoffe: Kleinstbausteine mit großer Wirkung*. Zuletzt abgerufen am 11. September 2021 URL: https://www.wissenschaft.de/gesundheit-medizin/mikronaehrstoffe-kleinstbausteine-mit-grosser-wirkung/

[L31] LanguageTooler GmbH: *openthesaurus*: Synonyme und Assoziationen. Zuletzt abgerufen am 09. Mai 2021 URL: https://www.openthesaurus.de/

[L32] Lebensmittelverband Deutschland e. V. (2017). *Liste der Zusatzstoffe und E-Nummern.* Zuletzt abgerufen am 11. Juli 2021 URL: https://www.lebensmittelverband.de/de/lebensmittel/inhaltsstoffe/zusatzstoffe/liste-lebensmittelzusatzstoffe-e-nummern

[L33] Luber, S., & Litzel, N. (2018). *Was ist ETL (Extract, Transform, Load)?* Zuletzt abgerufen am 24. September 2021 URL: https://www.bigdata-insider.de/was-ist-etl-extract-transform-load-a-776549/

[L34] Max Rubner-Institut. *Bundeslebensmittelschlüssel*: Was ist der BLS? Zuletzt abgerufen am 29. August 2021 URL: https://blsdb.de/bls?background

[L35] Max Rubner-Institut. *Bundeslebensmittelschlüssel*. Zuletzt abgerufen am 29. August 2021 URL: https://www.mri.bund.de/de/service/datenbanken/bundeslebensmittelschluessel/

[L36] MaxMind, Inc. *MAXMIND*: GeoLite2 Free Geolocation Data. Zuletzt abgerufen am 11. September 2021 URL: https://dev.maxmind.com/geoip/geolite2-free-geolocation-data

[L37] McMahon, P., Chiorean, M.-L., Coleman, S., & Askoolum, A. (2018). *The Guardian*: Bye bye Mongo, Hello Postgres. Zuletzt abgerufen am 22. Oktober 2021 URL: https://www.theguardian.com/info/2018/nov/30/bye-bye-mongo-hello-postgres

[L38] My Daily Bits LLC. (2020). Google Play Store: *Ernährungstagebuch*. Zuletzt abgerufen am 28. August 2021 URL: https://play.google.com/store/apps/details?id=com.dailybits.foodjournal&hl=de&gl=US

[L39] Neuleben, I. (2018). *Dokumentationspflicht und Aufbewahrungsfristen.* Zuletzt abgerufen am 18. Juli 2020 URL: Kassenärztliche Vereinigung Nordrhein: https://www.kvno.de/10praxis/30honorarundrecht/30recht/20dokupflicht/15_05_aufbewahrungsfristen/index.html

[L40] NIST. (2017). von COSINE DISTANCE: *Statistical Data Engineering Division Data plot*. Zuletzt abgerufen am 31. Mai 2021 URL: https://www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/cosdist.htm

[L41] Open Knowledge Foundation. *Open Data Commons*: Legal tools for Open Data Zuletzt abgerufen am 11. September 2021 URL: https://opendatacommons.org/licenses/odbl/

[L42] Paulsen, N., & Schenk, A. (2017). *Fast jeder Zweite nutzt Gesundheits-Apps. Bitkom e.V.* Zuletzt abgerufen am 11. Juni 2020 URL: https://www.bitkom.org/Presse/Presseinformation/Fast-jeder-Zweite-nutzt-Gesundheits-Apps.html

[L43] Presse- und Informationsamt der Bundesregierung. (2016). *Nährwertkennzeichnung ist Pflicht.* Zuletzt abgerufen am 11. Juli 2021 URL: https://www.bundesregierung.de/breg-de/aktuelles/naehrwertkennzeichnung-ist-pflicht-348186#:~:text=Verbraucherschutz%20N%C3%A4hrwertkennzeichnung%20ist%20Pflicht&text=Die%20N%C3%A4hrwerte%20eines%20Lebensmittels%20m%C3%BCssen%20in%20Tabellenform%20angegeb

[L44] Schneider, M. (2020). Analytics. TimescaleDB vs. influxDB: Zeitreihendatenbanken für das IIoT. Zuletzt abgerufen am 22. Oktober 2021 URL: https://www.inovex.de/de/blog/timescaledb-vs-influxdb-zeitreihen-iiot/

[L45] Seeger, A. (2021). *Ist das Falt-Smartphone bereit für den Massenmarkt?* Zuletzt abgerufen am 15. August 2021 URL: https://www.connect.de/testbericht/samsung-galaxy-z-fold-3-z-flip-3-erster-test-3201758.html

[L46] snoopmedia GmbH. *das ist drin: gemeinsam besser leben.* Zuletzt abgerufen am 11. September 2021 URL: http://das-ist-drin.de/

[L47] Snowflake Inc. *Schätzung der Ähnlichkeit von zwei oder mehr Sets*. Zuletzt abgerufen am 15. August 2021 URL: https://docs.snowflake.com/de/user-guide/querying-approximate-similarity.html

[L48] Solnichkin, A. (2018). Medium: *4 Best Time Series Databases To Watch in 2019*. Zuletzt abgerufen am 22. Oktober 2021 URL: https://medium.com/schkn/4-best-time-series-databases-to-watch-in-2019-ef1e89a72377

[L49] The Apache Software Foundation. Apache Spark: *RDD Programming Guide*. Zuletzt abgerufen am 13. September 2021 URL: https://spark.apache.org/docs/latest/rdd-programming-guide.html

[L50] The Apache Software Foundation. *Apache Spark*: Unified engine for large-scale data analytics. Zuletzt abgerufen am 22. Oktober 2021 URL: https://spark.apache.org/

[L51] The IEEE and The Open Group. (2004). *The Open Group Base Specifications Issue 6; 9. Regular Expressions*. Zuletzt abgerufen am 4. September 2020 URL: https://pubs.opengroup.org/onlinepubs/009695399/basedefs/xbd_chap09.html#tag_09_03_05

[L52] The Open Food Facts Team. *Open Food Facts - World*. Zuletzt abgerufen am 11. September 2021 URL: https://world.openfoodfacts.org/

[L53] THE OPEN FOOD REPO. *Eine offene community Datenbank für barcodierte Lebensmittel.* (Digital Epidemiology Laboratory (EPFL). Zuletzt abgerufen am 11. September 2021 URL: https://www.foodrepo.org/

[L54] the pandas development team. *pandas.DataFrame*. Zuletzt abgerufen am 13. September 2021 URL: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html

[L55] Timescale Inc. *PostgreSQL for time-series*. Zuletzt abgerufen am 13. September 2021 URL: https://www.timescale.com/

[L56] Timescale, Inc. *TimescaleDB*: How TimescaleDB works. Zuletzt abgerufen am 13. September 2021 URL: https://www.timescale.com/products/#How-it-works

[L57] TÜV Rheinland AG. (2021). *CE-Kennzeichnung*. Zuletzt abgerufen am 27. 07 2021 URL: https://www.tuv.com/germany/de/ce-kennzeichnung.html

[L58] TÜV SÜD AG. Von Risikomanagement nach EN ISO 14971. Zuletzt abgerufen am 24. Oktober 2021. URL: https://www.tuvsud.com/de-de/branchen/gesundheit-und-medizintechnik/risikomanagement-en-iso-14971

[L59] Verbraucherzentrale NRW e.V. (2020). *Lebensmittel-Kennzeichnung: Was muss drauf stehen?* Zuletzt abgerufen am 11. Juli 2021 URL: https://www.verbraucherzentrale.de/wissen/lebensmittel/kennzeichnung-und-inhaltsstoffe/lebensmittelkennzeichnung-was-muss-drauf-stehen-5430

[L60] *VERORDNUNG (EU) Nr. 1169/2011 DES EUROPÄISCHEN PARLAMENTS UND DES RATES.* (2011). Zuletzt abgerufen am 11. Juli 2021 URL: https://eur-lex.europa.eu/legal-content/DE/TXT/HTML/?uri=CELEX:02011R1169-20180101&from=DE

[L61] *VictoriaMetrics*. Zuletzt abgerufen am 22. Oktober 2021 URL: https://docs.victoriametrics.com/Single-server-VictoriaMetrics.html

[L62] W3Schools. *Python Casting*. Zuletzt abgerufen am 13. September 2021 URL: w3schools.com: https://www.w3schools.com/python/python_casting.asp

[L63] YAZIO GmbH. (2021). Yazio: *Willkommen zu einem gesünderen Leben.* Zuletzt abgerufen am 28. August 2021 URL: https://www.yazio.com/de

Abkürzungsverzeichnis

5G	Fünfte Generation (des Mobilfunkstandards)
ACC	Accuracy
AOP	Aspektorientierte Programmierung
API	Application Programming Interface
BI	Business Intelligence
BLS	Bundeslebensmittelschlüssel
BMBF	Bundesministerium für Bildung und Forschung
CE	Conformitè Europëenne
CRISP-DM	Cross Industry Standard Process for Data Mining Projects
CRP	Centre de Recherche Public
CSV	Comma-separated Values
DAAB	Deutscher Allergie- und Asthmabund
DB	Datenbank
DG	Gesamtplausibilitätswert der Plausibilitätsanalyse der Tagebuchdatensätze
DI	Plausibilitätswert von Inhaltstofflisten der Plausibilitätsanalyse der Tagebuchdatensätze
DiDiER	Digitale Dienstleistungen in der Ernährungsberatung
DIN	Deutsches Institut für Normung
DN	Plausibilitätswert von Nährwerten der Plausibilitätsanalyse der Tagebuchdatensätze
EAN	European Article Number
EN	Europäische Norm
EPFL	École Polytechnique Fédérale de Lausanne
ERM	Entity Relationship Model
ETL	Extract, Transform and Load
EU	Europäische Union
FCDB	Food Composition Database
FDWH	Food Data Warehouse
FN	Falsch-Negativ-Wert

FP	Falsch-Positiv-Wert
FPR	Falsch-Positiv-Rate
FPDB	Food Product Database
GB	Gigabyte
GHz	Gigahertz
GTIN	Global Trade Item Number
ID	Identifier
IKT	Informations- und Kommunikations-Technologie
INI	Initial (Dateiendung für Konfigurationsdateien)
ISO	International Organization for Standardization
JSON	JavaScript Object Notation
KNIME	Konstanz Information Miner
LIST	Luxembourg Institute of Science and Technology
MDR	Medical Device Regulation
ML	Machine Learning
MRI	Max Rubner-Institut
MVC	Model View Controller
RAM	Random Access Memory
RDD	Resilient Distributed Dataset
REST	Representational State Transfer
RN	Richtig-Negativ-Wert
ROC	Receiver Operating Characteristics
RP	Richtig-Positiv-Wert
RPR	Richtig-Positiv-Rate
PCA	Principal Component Analysis
PREC	Precision
SBLS	BLS-Datensatzschlüssel
SHA	Secure Hash Algorithm
SI	Système international d'unités
SOUNDEX	Soundexpression-Algorithmus
SQL	Structured Query Language
TG	Gesamtplausibilitätswert der Plausibilitätsanalyse der Testdatensätze
TI	Plausibilitätswert von Inhaltstofflisten der Plausibilitätsanalyse der Testdatensätze

TN	Plausibilitätswert von Nährwerten der Plausibilitätsanalyse der Testdatensätze
TOS-DQ	Talend Open Studio for Data Quality
XML	Extensible Markup Language
WiFi	Wireless Fidelity

Abbildungsverzeichnis

Abbildung 1.1 *CRISP-DM-Modell (in Anlehnung an Chapman et al. (2000) [15] und Cleve & Lämmel (2020) [18])* 15
Abbildung 2.1 *Beispiel eines eindimensionalen Barcodes des Produktes Steinofen-Pizza der Marke Wagner* 24
Abbildung 2.2 *Beispiel einer Inhaltsstoffliste des Produktes Steinofen-Pizza der Marke Wagner* 24
Abbildung 2.3 *Beispiel einer Nährwerteliste des Produktes Steinofen-Pizza der Marke Wagner* 24
Abbildung 2.4 *Auszug aus ERM der WikiFood-Datenbank* 30
Abbildung 2.5 *MongoDB von OpenFoodFacts.org: Auszug der Attributfelder* 32
Abbildung 2.6 *Auszug einer Produktdatenselektierung auf der das-ist-drin.de Webseite* 34
Abbildung 2.7 *Durchgeführte Schritte während des Clusterings (in Anlehnung an Wissuwa et al. (2005) [75])* 44
Abbildung 2.8 *ROC Diagramm mit Diagonalen und Beschreibung der Bereiche* 49
Abbildung 3.1 *Beispiel von Qualitätsmängeln zweier Auszüge unterschiedlicher Datenquellen mit Lebensmittelproduktdaten* 53
Abbildung 3.2 *Ontologie-Modell der Datenelemente des FDWH* 55
Abbildung 3.3 *Beispiel einer Zeichenkette-Musteranalyse in TOS-DQ* 60
Abbildung 3.4 *Grafische Darstellung des Snowflake-Schema* 64
Abbildung 3.5 *Darstellung des gesamten FDWH-Schema* 65
Abbildung 3.6 *Auszug aus der FDWH-Datenbank als Beispiel für die ID-Änderung bei gleichzeitiger Änderung in der Zusammensetzung des Lebensmittelproduktes* 67
Abbildung 3.7 *Beispiel der Struktur eines Dataframes im Pandas-Framework* 68
Abbildung 3.8 *Grafische Darstellung des ETL-Prozesses* 70
Abbildung 3.9 *Beispiel der Ähnlichkeitsanalyse zweier Produktbezeichnungen* 75
Abbildung 3.10 *Darstellung des Verhältnisses von Kohlenhydraten zu Fett der BLS-Datensätze als Punktdiagramm* 77
Abbildung 3.11 *Darstellung des Verhältnisses von Kohlenhydraten zu Proteinen der BLS-Datensätze als Punktdiagramm* 78
Abbildung 3.12 *Darstellung der aus den BLS-Datensätzen gebildeten Clustern nach der PCA-Dimensionsreduzierung* 79
Abbildung 3.13 *Wortwolke aus den Kategorien der BLS-Datensätze, die dem Cluster cluster_16 zugeordnet sind* 80
Abbildung 3.14 *Entscheidungsbaum zur Bestimmung der Plausibilität der Nährwerte* 80

Abbildung 3.15 *Entscheidungsbaum zur Bestimmung der Plausibilität der Inhaltsstoffe* 83
Abbildung 3.16 *Beispiel einer FDWH-API-Anfrage für eine Produktübersicht im JSON-Format* 85
Abbildung 3.17 *Beispiel einer FDWH-API-Anfrage für Details eines Produktes im JSON-Format* 86
Abbildung 3.18 *Weboberfläche für Testabfragen innerhalb des FDWH* 87
Abbildung 3.19 *Darstellung der FDWH-API-Anbindung an das Spark Framework und dessen Bearbeitung der Daten im Hauptspeicher sowie an die App des Anwenders* 91
Abbildung 3.20 *Anzahl der Gesamtattributwerte vor und nach der Bereinigung* 91
Abbildung 3.21 *ROC Diagramm mit RPR- und FPR-Daten der Validierung die sich links über der Diagonalen befinden* 96

Tabellenverzeichnis

Tabelle 2.1 *Gesundheitliche Beeinträchtigungen und Krankheitssymptome in Abhängigkeit mit Lebensmittelbestandteilen* *21*
Tabelle 2.2 *Beschreibung der Stellen im SBLS* *29*
Tabelle 2.3 *Für das FDWH relevante Attribute mit Bezug zu den Datenquellen, in denen sie enthalten sind* *38*
Tabelle 2.4 *Wahrheitsmatrix zur Berechnung von Genauigkeit, Präzision, RPR und FPR* *49*
Tabelle 3.1 *Attributfelder der FDWH-Datenbank inklusive Datentyp Zuordnung* *58*
Tabelle 3.2 *Ermittelte reguläre Ausdrücke der Speicherung von Zeichenketten in den externen Datenquellen* *61*
Tabelle 3.3 *Ermittelte reguläre Ausdrücke für die einheitliche Speicherung von Zeichenketten im FDWH* *62*
Tabelle 3.4 *Beschreibung der regulären Ausdrücke aus Tabelle 3.3* *63*
Tabelle 3.5 *Informationen zur Generierung der FDWH-IDs* *66*
Tabelle 3.6 *Anzahl Attributwerte der jeweiligen Datenquellen und insgesamt, vor und nach der Bereinigung* *90*
Tabelle 3.7 *Wahrheitsmatrix der Plausibilitätsanalyse der Tagebuchdatensätze im Bereich der Nährwerteattribute* *95*
Tabelle 3.8 *Wahrheitsmatrix der Plausibilitätsanalyse der Tagebuchdatensätze im Bereich der Inhaltsstoffattribute* *96*
Tabelle 3.9 *Wahrheitsmatrix der Gesamtplausibilitätsanalyse der Tagebuchdatensätze* *96*
Tabelle 3.10 *Wahrheitsmatrix der Plausibilitätsanalyse der 100 Testdatensätze im Bereich der Nährwertetattribute* *96*
Tabelle 3.11 *Wahrheitsmatrix der Plausibilitätsanalyse der 100 Testdatensätze im Bereich der Inhaltsstoffattribute* *96*
Tabelle 3.12 *Wahrheitsmatrix der Gesamtplausibilitätsanalyse der 100 Testdatensätze* *97*
Tabelle 4.1 *Beispiele von Daten für eine Anwendung der Ähnlichkeits- und Plausibilitätsanalyse* *105*

Anhang

Anhang 1:

Tabelle mit Details zum durchgeführten Test der Gesundheitsapps (siehe Abschnitt 2.1).

App-Name	Ent-wickler	Zugriff auf FCDB	Zugriff auf FPDB	Qualität FCDB-Daten	Qualität FPDB-Daten	Gesamt-qualität der DB
Ernährungs-tagebuch	My Daily Bits LLC	ja	ja	+	+	+
Food Diary	WeCode?	ja	nein	-	/	/
Food Diary and Journal	Random Apps Inc.	nein	nein	/	/	/
See How You Eat Food Diary App	Health Revolution Ltd.	nein	nein	/	/	/
DigestIT Ernährungs-tagebuch	Enrico Wegner	nein	nein	/	/	/
My Symptoms Food Diary und Symptom Tracker	n Sky Gazer Labs Ltd.	ja	ja	+	-	0
IEAT Well	Premium Health& Fitness Apps	nein	nein	/	/	/
YAZIO Kalorien-zähler	YAZIO	ja	ja	++	+	++
Ernährungs-tagebuch deluxe	IT Service Herzog	ja	ja	-	-	-

Recipes & Nutrition	Edaman	nein	nein	/	/	/
Calorie Counter MyNetDiary	MyNet Diary.com	ja	ja	-	-	-
Ernährungs-tagebuch	Perfect Tools	nein	nein	/	/	/
Intolerance Food Diary	Alder Agarik	nein	nein	/	/	/
Diet Diary	Can Yapan	nein	nein	/	/	/
Codecheck Lebensmittel und Cosmetic Scanner	Codecheck AG	nein	ja	/	++	/
Lebensmittel Intoleranz Liste	cr3ative. info	ja	nein	+	/	/

Legende:

++	Sehr gute Qualität
+	Gute Qualität
0	Mittelmäßige Qualität
-	Schlechte Qualität
--	Sehr schlechte Qualität
/	Qualität nicht ermittelbar

Anhang 2:

Auszug des Ergebnisses einer Produktübersichtanfrage der FDWH-API, mit Bezug zu der in Abbildung 3.16 dargestellten JSON-Anfrage:


```
{
  "productOverview": {
    "hits": 10,
    "page": 1,
    "product": [
      {
        "id": 970511,
        "name": "Milchreis Zimt",
        "productDescriptions": {
          "description": [
            "Kunststoff"
          ]
        },
        "brand": {
          "id": "53cd2da0156b3c1a5258ae12bc6b5897052e709d8a03e3dbfc2ebdf9b599cd41",
          "name": "Desira"
        },
        "origin": {
          "id": "af6e56aacd8097c3fb7597b08ecb3d87a040f6ab58067a4b8ff9f5b747adb9be",
          "name": "Deutschland"
        },
        "eans": {
          "ean": [
            "4061458023115"
          ]
        },
        "categories": {
          "categories": [
            {
              "id": "1.19.7",
              "name": "Essen > Getreidewaren > Reis"
            }
          ]
        },
        "insertDate": "2019-18-19 07:18:20",
        "modificationDate": "2019-18-19 07:18:20"
      },
      {
        "id": 976769,
        "name": "Milchreis Zimt",
        "productDescriptions": {
          "description": [
            null
          ]
        },
        "brand": {
          "id": "4595df3507ca67ef98c25fbb9cd9537d9326d6549fdb596abf433a474e5a6a9f",
          "name": "Milbona"
        },
        "origin": {
          "id": "af6e56aacd8097c3fb7597b08ecb3d87a040f6ab58067a4b8ff9f5b747adb9be",
          "name": "Deutschland"
        },
        "eans": {
          "ean": [
            "20277680"
          ]
        },
        "categories": {
          ...
        "insertDate": "2019-40-25 11:40:11",
        "modificationDate": "2019-40-25 11:40:11"
      }
    ]
  }
}
```

Anhang 3:

Auszug des Ergebnisses einer Produktdetailanfrage der FDWH-API, mit Bezug zu der in Abbildung 3.17 dargestellten JSON-Anfrage:


```
{
  "productDetails": {
    "product": [
      {
        "id": 970511,
        "name": "Milchreis Zimt",
        "productDescriptions": null,
        "brand": {
          "id": "53cd2da0156b3c1a5258ae12bc6b5897052e709d8a03e3dbfc2ebdf9b599cd41",
          "name": "Desira"
        },
        "origin": {
          "id": "af6e56aacd8097c3fb7597b08ecb3d87a040f6ab58067a4b8ff9f5b747adb9be",
          "name": "Deutschland"
        },
        "eans": {
          "ean": [
            "4061458023115"
          ]
        },
        "categories": {
          "categories": [
            {
              "id": "1.19.7",
              "name": "Essen > Getreidewaren > Reis"
            }
          ]
        },
        "insertDate": "2019-18-19 07:18:20",
        "modificationDate": "2019-18-19 07:18:20",
        "contents": {
          "contents": [
            {
              "value": "200.0",
              "unit": "g"
            }
          ]
        },
        "ingredientLists": {
          "ingredLists": [
            {
              "list": "Vollmilch, Buttermilch, Wasser, Zucker, 8,5 % Reis, Zimt, Stärke, ...
              "lang": "de",
              "adopted": false,
              "focusRelation": null,
              "plausible": 0
            }
          ]
        },
        "mainAllergens": {
          "mainAllergens": [
            {
            {
            {
            {
              "name": "kr",
              "contained": "k"
            },
            {
              "name": "la",
              "contained": "j"
            },
```

```
        ]
      },
      "nutritionFacts": {
        "nutritionFacts": [
          {
            "name": "Kohlenhydrate in mg pro 100g",
            "nutrTyp": "carbohydrates",
            "value": 20700,
            "propValue": 100,
            "siUnit": "mg",
            "propUnit": "g"
          },
          {
            "name": "Fett in mg pro 100g",
            "nutrTyp": "fat",
            "value": 1300,
            "propValue": 100,
            "siUnit": "mg",
            "propUnit": "g"
          },
          {
            "name": "Proteine in mg pro 100g",
            "nutrTyp": "proteins",
            "value": 2900,
            "propValue": 100,
            "siUnit": "mg",
            "propUnit": "g"
          },
          {
            "name": "Energy in kcal pro 100g",
            "nutrTyp": "energy_kcal",
            "value": 113,
            "propValue": 100,
            "siUnit": "kcal",
            "propUnit": "g"
          }
        ],
        "adopted": false,
        "focusRelation": null,
        "plausible": 0,
        "cluster": ""
      },
      "similarFoods": {
        "similarFoods": null,
        "lastTotalSimilVal": 0
      },
      "ingredientContextRef": [
        true
      ],
      "nutritionContextRef": [
        true
      ],
      "allergenContextRef": [
        true
      ]
    }
  ]
},
"allCategories": null
}
```