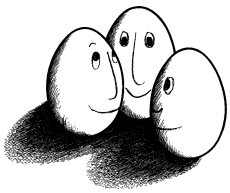
Diplomarbeit

Analyse von Krankenversichertendaten zur Identifikation psychischer Krankheiten

Alice Czerniejewski



Diplomarbeit
am Fachbereich Informatik
der Universität Dortmund

Dortmund, 15. Juli 2008

Betreuer:

Prof. Dr. Katharina Morik
Dipl.-Inform. Ingo Mierswa

Danksagung

Ich bedanke mich bei Prof. Dr. Katharina Morik und Dipl.-Inform. Ingo Mierswa für die Betreuung dieser Arbeit, sowie bei allen anderen Personen, die mich während dessen unterstützt haben.

Inhaltsverzeichnis

Abkürzungsverzeichnis

Anz.	Anzahl
AOK	Allgemeine Ortskrankenkasse
AOP	Ambulantes Operieren
AU	Arbeitsunfähigkeit
BKK	Betriebskrankenkasse
CRISP-DM	Cross Industry Standard Process Modell for Data Mining
DAK	Deutsche Angestellten-Krankenkasse
DIMDI	Deutsches Institut für Medizinische Dokumentation und Information
eGK	elektronische Gesundheitskarte
GEK	Gmünder ErsatzKasse
GKV	Gesetzliche Krankenversicherung
HEK	Hanseatische Krankenkasse
HZK	Hamburgische Zimmererkrankenkasse
ICD	Internationale statische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme
IKK	Innungskrankenkassen
KDD	Knowledge Discovery in Databases
KEH	Krankenkasse Eintracht Heusenstamm
KKH	Krankenhaus
KV	Kassenärztliche Vereinigung
KVNR	Krankenversicherungsnummer
KZV	Kassenzahnärztliche Vereinigung
PVS	Praxis-Verwaltungs-Softwaresystem
RVNR	Rentenversicherungsnummer
SGB	Sozialgesetzbuch
VSD	Versichertenstammdaten
WHO	Weltgesundheitsorganisation

Abbildungsverzeichnis

Tabellenverzeichnis

1. Einleitung

In einem ersten Schritt wird zunächst die Problemstellung der Arbeit thematisiert, um daraus die Ziele ableiten zu können. Des Weiteren wird die Gliederung der Arbeit aufgezeigt.

1.1. Problemstellung und Ziele der Arbeit

Der Leistungskatalog der gesetzlichen Krankenversicherung beinhaltet u.a. Vorsorge, Behandlung von Krankheiten, Empfängnisverhütung, medizinische Rehabilitation, sowie Leistungen zur Prävention von Krankheiten. Die Tendenz der letzten 20 Jahre hat gezeigt, dass durch verbesserte medizinische Behandlungen und Entwicklung von Präventionsprogrammen die Dauer der Arbeitsunfähigkeit aufgrund bestimmter Krankheiten deutlich gesenkt werden konnte. Ein gutes Beispiel dafür sind Krankheiten des Muskel-Skelett-Systems. Durch Rückenschulen und Aufklärung konnte die Arbeitsunfähigkeitsdauer der Versicherten halbiert werden.

Der Bereich der psychischen Störungen ist leider von dieser Entwicklung ausgeschlossen, siehe Abbildung 1.1.

Abbildung 1.1.: Krankheitsarten - Trends seit 1976 nach [BKK05]

In den letzten 20 Jahren haben sich die Arbeitsunfähigkeitsmeldungen der Versicherten aufgrund psychischer Störungen fast verdoppelt. Diese hohen Arbeitsunfähigkeitszeiten

verursachen einen Verlust von produktiven Lebensjahren und damit eine Verminderung der Arbeitsproduktivität.

Besonders besorgniserregend sind die Ergebnisse des Bundes-Gesundheitssurveys von 1998/99, wonach jeder dritte Bundesbürger (32, 1%) im Laufe seines Lebens einmal an einer pychischen Störung erkranken wird, aber nur jeder Dritte der Betroffenen aus diesem Grund mindestens einmal professionelle Hilfe in Anspruch nehmen wird (vgl. [JW02] und [JKW04]).

Es wird mit Hochdruck daran gearbeitet entsprechende Präventionsprogramme zu entwickeln, die zum einen die Lebensqualität der Versicherten steigern und zum anderen die Kosten der Krankenversicherung senken sollen. Dies ist auch im Sinne der Versicherten, da die Krankenkassen durch die Beiträge ihre Kosten an die Versicherten weitergeben. Dazu müssen frühzeitig die Personen identifiziert werden, welche gefährdet sind an psychischen Störungen zu erkranken, damit ihnen im Vorfeld ein Präventionsprogramm angeboten werden kann.

Diverse Studien, wie zum Beispiel [BKK05], [TK005], [DAK05] oder [AOK05], beschäftigten sich bereits mit der Ursachenforschung. Die Ursachen für psychische Erkrankungen sind sehr komplex, doch einige Einflussfaktoren haben sich schon herauskristallisiert.

Die Häufigkeit und Dauer der Arbeitsunfähigkeitsmeldungen wegen psychischer Erkrankungen unterscheidet sich u.a. nach dem Geschlecht. Frauen sind durchschnittlich einen halben Tage länger und doppelt so häufig wie Männer arbeitsunfähig gemeldet.

Auffällig sind auch die Altersverläufe bestimmter psychisch Kranker. Im Alter steigt die Dauer der Arbeitsunfähigkeitsmeldungen aufgrund von affektiven Störungen, wie Depressionen, neurotischen und Belastungsstörungen deutlich an. So sind Frauen im Alter von 60 Jahren durchschnittlich 1, 3 Tage arbeitsunfähig gemeldet und Männer 0, 6 Tage.

Zusätzlich haben Faktoren, wie Beruf und Lebenssituation, einen entscheidenden Einfluss auf die Krankheitsentwicklung. Tabellen 1.1 und 1.2 verdeutlichen die Bedeutung der Berufswahl auf die Dauer der Krankheitszeiten.

So sind Beschäftigte in der Krankenpflege ebenso wie Sozialarbeiter besonders gefährdet. Das größte Risiko psychisch zu erkranken liegt jedoch bei den Arbeitssuchenden. So sind arbeitssuchende Männer durchschnittlich 1, 92 Tage im Jahr krank gemeldet und Frauen sogar 2, 75 Tage.

Die bisherigen Studien beschäftigten sich mit der Erforschung der Zusammenhänge zwischen den sozialen Merkmalen, wie Alter, Beruf, Geschlecht, Wohnort und der diagnostizierten psychischen Erkrankung. Die Bedeutung der Vorerkrankungen eines psychisch Kranken wurden noch nicht betrachtet.

Das Ziel dieser Diplomarbeit ist die Indentifizierung von Mustern in den Krankengeschichten eines psychisch Kranken, um Patienten, die erst einen Teil einer typischen Krankengeschichte durchlaufen haben, einen Therapievorschlag und ein Präventionsprogramm anbieten zu können.

Beruf	AU [Tage/Jahr]
Hilfsarbeiter	1,17
Postverteiler	1,24
Straßenreiniger	1,24
Fernmeldemonteure und - handwerker	1,30
Helfer in der Krankenpflege	1,33
Eisenbahnbetriebsregler und -schaffner	1,53
Heimleiter, Sozialpädagogen	1,54
Wächter, Aufseher	1,68
Sozialarbeiter, Sozialpfleger	1,83
Krankenpfleger	1,87
Arbeitslose	1,92

Tabelle 1.1.: Krankheitszeiten der Männer [BKK05]

Beruf	AU [Tage/Jahr]
Köchinnen	1,83
Elektrogerätemontiererinnen	1,89
Hauswirtschaftliche Betreuerin	1,92
Sonstige Technikerin	2,06
Heimleiterinnen und Sozialpädagoginnen	2,08
Sonstige Montiererinnen	2,19
Hilfsarbeiterinnen	2,32
Sozialarbeiterinnen und Sozialpflegerinnen	2,47
Telefonistinnen	2,5
Helferinnen in der Krankenpflege	2,78
Arbeitslose	2,75

Tabelle 1.2.: Krankheitszeiten der Frauen [BKK05]

1.2. Gliederung der Arbeit

Zunächst wird in Kapitel 2 der Begriff der Wissensentdeckung in Datenbanken (KDD) definiert um daran anschließend die einzelnen Phasen des KDD-Prozesses anhand des CRISP-DM [CCK+00] näher zu erläutern.

Um einen Einblick in das sehr komplexe Sachgebiet der gesetzlichen Krankenversicherung zu bekommen, werden in Kapitel 3 die Grundsätze, die Aufgaben und die beteiligten Organisationen der gesetzlichen Krankenversicherung, sowie ihre Beziehungen untereinander beschrieben. Ebenso werden in diesem Kapitel die Herkunft, der Umfang und die Datenqualität der Krankenkassendaten behandelt.

Eine wichtige Rolle in dem KDD-Prozess und in dieser Arbeit spielt die Datenaufbereitung, welcher das Kapitel 4 gewidmet ist. Neben der Datenbereinigung und -selektion werden in diesem Kapitel fünf verschiedene Möglichkeiten, welche die Krankengeschichte eines Versicherten repräsentieren, vorgestellt.

Kapitel 5 behandelt die Lernverfahren, die im Rahmen dieser Diplomarbeit zur Mustererkennung in Krankengeschichten wurden. Darunter zählen Naive Bayes [HK06], Perceptron [MP88], Subgruppenentdeckung [Sch05] und GSP [AS95]. Anschließend werden in Kapitel 6 die mit den zuvor beschriebenen Lernverfahren erzielten Ergebnisse dargestellt.

In Kapitel 7 werden abschließend die Ergebnisse dieser Arbeit zusammengefasst und es wird ein Ausblick auf zukünftige Arbeiten und Entwicklungen gegeben.

2. Wissensentdeckung in Datenbanken am Beispiel des Prozessmodells CRISP-DM

In diesem Kapitel wird der Begriff der Wissensentdeckung in Datenbanken definiert und der Prozess der Wissensentdeckung anhand des CRISP-DM-Modells veranschaulicht.

2.1. Wissensentdeckung in Datenbanken (KDD)

Folgender Satz aus [FPSS96] wird häufig in wissenschaftlichen Arbeiten zitiert, um den Begriff der Wissensentdeckung in Datenbanken bzw. Knowledge Discovery in Databases (KDD) zu definieren:

Definition 2.1 (Wissensentdeckung in Datenbanken) *Wissensentdeckung in Datenbanken ist der nichttriviale Prozess der Identifikation gültiger, neuer, potentiell und verständlicher Muster in (großen) Datenbeständen.*

2.2. CRISP-DM

In der Forschung und in der Praxis wurden verschiedene Prozessmodelle für die Wissensentdeckung ausgearbeitet. Im Folgenden wird der Prozess der Wissensentdeckung anhand des CRISP-DM (Cross Industry Standard Process Modell for Data Mining nach [CCK+00] erläutert. Dieses Modell teilt den Prozess der Wissensentdeckung in sechs Phasen auf:

Anwendungsanalyse In dieser initialen Phase soll zunächst das Anwendungsgebiet analysiert werden, um darauf basierend die Projektziele formulieren zu können. Diese umfasst die Abgrenzung des Sachgebiets, Analyse des Sachgebiets inkl. Recherche nach bereits geleisteten Vorarbeiten (Expertenwissen, Studien etc.), Formulierung einer klaren Fragestellung, die von einem Data Mining-Verfahren gelöst werden soll, sowie die Erstellung eines Projektplans und Auswahl geeigneter Werkzeuge.

Domänenanalyse Ziel der zweiten Phase ist, das Potential der zur Verfügung stehender Daten zu erforschen. Dies umfasst eine Bestandsaufnahme der vorhandenen Datenquellen, Sichtung der Daten, Erstellung einer Datensatzbeschreibung, Analyse der Attribute hinsichtlich Bedeutung und Zusammenhängen, sowie die Analyse der Datenqualität und Identifizierung von Fehlerquellen.

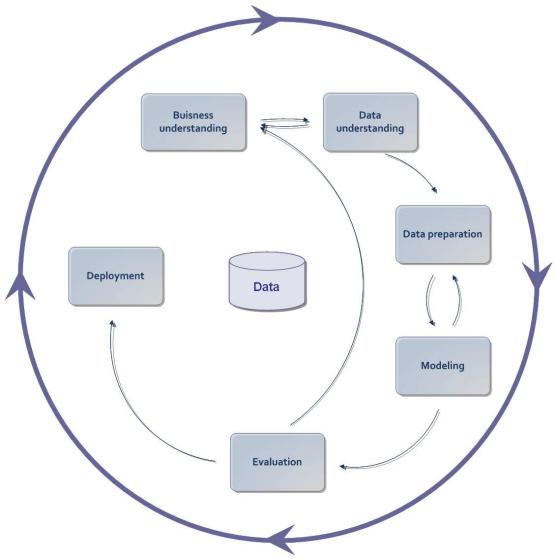


Abbildung 2.1.: Phasen des CRISP-DM

Datenaufbereitung In dieser Phase wird aus den Rohdaten die Eingabe für die Data Mining-Verfahren erzeugt, welche die zuvor formulierte Fragestellung lösen könnten. Sie umfasst die Selektion geeigneter Tabellen, Attribute und Datensätze, die Ableitung höherwertiger oder andersartiger Attribute aus den ursprünglichen, sowie die notwendige Formattransformation für den in der nächsten Phase eingesetzten Data-Mining-Algorithmus.

Data Mining Hier findet die eigentliche Wissensentdeckung statt. Von den bisher gewonnenen Erkenntnissen hängt die Auswahl des geeigneten Data Mining-Verfahrens ab, welches im Anschluss auf den zuvor selektierten und aufbereiteten Daten angewendet werden soll.

Evaluierung Die Evaluierungsphase sieht vor, alle bisher erzielten Projektergebnisse zu überprüfen, insbesondere die Qualität der Ergebnisse sowie den Abgleich mit den zuvor definierten Projektzielen.

Anwendungsphase Diese Phase hängt von den definierten Projektzielen ab. Konnte ein gut geeignetes Data Mining-Verfahren, welches die anfangs gestellte Fragestellung löst, identifiziert werden, kann je nach Projektziel ein Bericht oder eine Präsentation über die erzielten Ergebnisse ausreichen. Mit sehr viel mehr Aufwand ist der Transfer des Wissensentdeckungsprozesses in die operativen Geschäftsprozesse verbunden.

Jede der sechs Phasen ist in mehrere Teilschritte untergliedert, jede Phase bzw. jeder Teilschritt wird zumeist mehrfach durchlaufen. Nach Abschluss jeder Phase werden die Ergebnisse mit den Zielen abgeglichen und entsprechend die nächsten Schritte geplant. Dabei ist es durchaus üblich, zu bereits durchlaufenen Phasen zurückzukehren. Abbildung 2.1 veranschaulicht die Übergänge zwischen den einzelnen Phasen.

3. Die gesetzliche Krankenversicherung

Die gesetzliche Krankenversicherung (GKV) ist neben der Renten-, Pflege-, Arbeitslosen- und Unfallversicherung ein Zweig der Sozialversicherung in der Bundesrepublik Deutschland, deren Rechtsgrundlage das Sozialgesetzbuch V [SGB08a] ist. Der soziale Auftrag der GKV besteht darin, vollen Versicherungsschutz im Krankheitsfall unabhängig von der finanziellen Leistungsfähigkeit des einzelnen Versicherten zu gewährleisten. Der Versicherte erhält im Krankheistfall unmittelbar eine Versorgung, ohne gegenüber den Ärzten direkte Zahlungen leisten zu müssen.

Die gesetzliche Krankenversicherung ist eine Pflichtversicherung. Jeder Arbeitnehmer, dessen Arbeitsentgelt die Jahresarbeitsentgeldgrenze, die aktuell bei 47.250 € liegt, nicht übersteigt, muss sich bei einem Träger der gesetzlichen Krankenversicherung versichern. Übersteigt das monatliche Einkommen die Jahresarbeitsentgeltgrenze, entfällt die Pflichtversicherung und der Arbeitnehmer hat die Möglichkeit, sich entweder freiwillig weiter bei einem Träger der gesetzlichen Krankenversicherung zu versichern oder zu einer privaten Krankenversicherung zu wechseln. Die Mitgliedsbeiträge werden unabhängig vom individuellen Risiko, wie Alter, Geschlecht, Krankengeschichte etc. erhoben. Sie richten sich nach dem monatlichen Bruttoeinkommen des Versicherten. Familienangehörige eines gesetzlich Versicherten, die über kein eigenes Einkommen verfügen, werden beitragsfrei mitversichert.

3.1. Die gesetzlichen Krankenkassen

Träger der gesetzlichen Krankenversicherung sind die gesetzlichen Krankenkassen. Zurzeit gibt es sieben Kassenarten und etwa 200 Krankenkassen, vgl 3.1.

Anzahl	Krankenkassenart
7	Angestellten-Krankenkassen (Barmer, DAK, Techniker Krankenkasse, KKH, HEK)
3	Arbeiter-Ersatzkassen (GEK, HZK, KEH)
17	Allgemeine Ortskrankenkassen (AOK)
199	Betriebskrankenkassen (BKK)
1	Knappschaft
16	Innungskrankenkassen (IKK)
9	Landwirtschaftliche Krankenkassen

Tabelle 3.1.: Kassenarten

Die einzelnen Krankenkassen sind, je nach Organisationsprinzip, zu Verbänden auf Landes- und/oder Bundesebene zusammengeschlossen. Zu den Aufgaben der Verbände zählen u.a. die Vertretung der gemeinsamen Interessen im politischen Raum, Betreuung und Beratung der Mitgliedskassen bei der Durchführung ihrer Aufgaben, sowie Verhandlung und Abschluss von Verträgen mit den Leistungserbringern.

3.2. Leistungserbringer

Ärzte, Zahnärzte, Krankenhäuser und Apotheken, aber auch eine Vielzahl anderer Gesundheitsleistungsanbieter, wie Physiotherapeuten oder Hebammen stellen im Rahmen der gesetzlichen Krankenversicherung Leistungen bereit. Sie werden unter dem Begriff Leistungserbringer zusammengefasst. Leistungserbringer sind ebenfalls in Verbänden auf Bundes- oder Landesebene organisiert, welche die Interessen ihrer Mitglieder als Vertragspartner gegenüber den gesetzlichen Krankenkassen vertreten.

3.3. Das Beziehungsfünfeck

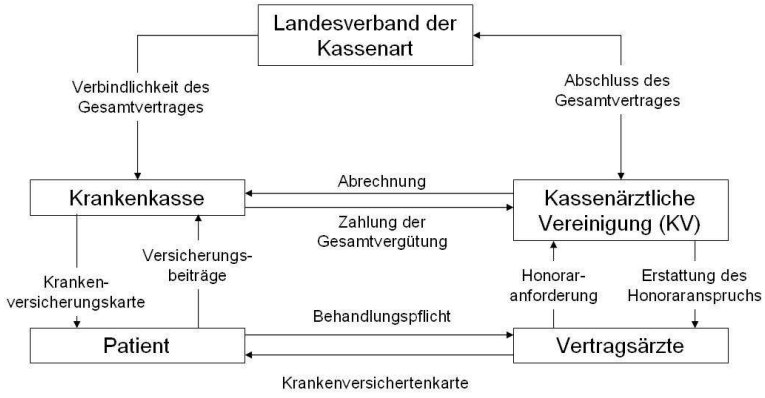


Abbildung 3.1.: Beziehungsfünfeck nach [Qua07]

Abbildung 3.1 veranschaulicht die komplexen Beziehungen zwischen Leistungserbringern, Patienten und gesetzlichen Krankenkassen. Die Krankenkassen sind, wie bereits beschrieben, auf Landes- bzw. Bundesebene zu Verbänden zusammmengeschlossen. Die Leistungserbringer sind ebenso zu Verbänden, u.a. kassenärztlichen beziehungsweise kassenzahnärztlichen Vereinigungen (KV bzw. KZV), zusammengeschlossen. Auf Landes- bzw. Bundesebene werden nun Verträge zwischen den gesetzlichen Krankenkassen und den Leistungserbringern geschlossen.

Der Versicherte einer Krankenkasse zahlt monatlich seine Beiträge und erhält eine Krankenversicherungskarte als Nachweis seiner Mitgliedschaft. Benötigt er im Krankheitsfall ärztliche Leistungen, kann er einen Vertragsarzt seiner Wahl aufsuchen. Durch Vorlage

der Krankheitsversichertenkarte belegt er sein Anrecht auf eine Behandlung.

Der Arzt führt für seinen Patienten eine oder mehrere Leistungen durch. Die Vergütung des Arztes wird von der Krankenkasse nicht direkt an ihn entrichtet. Er rechnet seine Leistungen gegenüber seiner kassenärtzlichen Vereinigung ab. Die KV bereitet die Daten auf, übermittelt die Einzelfallnachweise an die zuständige Krankenkasse und rechnet mit dieser ab oder erhält von ihr eine vereinbarte Gesamtvergütung. Diese wird von der KV entsprechend der abgerechneten Einzelleistungsnachsweisen an die Ärzte verteilt.

3.4. Krankenkassendaten

Dieses Kapitel beschäftigt sich mit der Datenherkunft, der Bedeutung der Attribute und deren Beziehungen, sowie der Datenqualität. An dieser Stelle sei angemerkt, dass die hier beschriebenen Daten von einem Rechenzentrum einer Krankenkasse vorselektiert bereitgestellt wurden und die Datenstruktur daher nicht den vertraglich und gesetzlich festgelegten Abrechnungsinformationen (vgl. [DAL08] und [SGB08a]) entspricht.

3.4.1. Versichertenstammdaten (VSD)

Eine Krankenkasse muss laut § 288, SGB V [SGB08a] ein Versichertenverzeichnis führen, das alle Angaben, die zur Feststellung der Versicherungspflicht bzw. - berechtigung, die zur Berechnung und Einziehung der Beiträge notwendig sind, enthält. Zu diesen Informationen zählen u.a. Name, Geschlecht, Alter, Arbeitgeber, Arbeitsverhältnis, Wohnort, sowie Renten- und Krankenversicherungsnummer.

Krankenversicherungsnummer

Die Krankenversicherungsnummer wird für jeden Versicherten kassenintern und nach eigenen Systematiken generiert. Diese Nummer enthält im Gegensatz zu der Rentenversicherungsnummer keinerlei personenbezogene Daten und ist nicht dauerhaft. Durch Veränderung des Versichertenverhätnisses (Arbeitslosigkeit, Eintritt in den Ruhestand oder Beginn eines Arbeitverhältnisses etc.) wird jedes Mal eine neue Krankenversicherungsnummer generiert. Sie wird zur Abrechnung der erbrachten Leistung durch die Leistungserbringer benötigt.

Rentenversicherungsnummer

Die Rentenversicherungsnummer ist in § 147, SGB VI [SGB08b] gesetzlich festgelegt und wird von den Trägern der Rentenversicherung vergeben. Sie wird erstmalig bei Aufnahme einer Beschäftigung, was mit der Eröffnung eines Rentenversicherungskontos einhergeht, generiert, ist eindeutig und gilt ein Leben lang. Die Generierung der zwölfstelligen Vericherungsnummer folgt einer gesetzlich vorgeschriebenen Systematik. Tabelle 3.2 veranschaulicht den Aufbau der Rentenversicherungsnummer.

Stellen	Bedeutung
1-2	Bereichsnummer des vergebenden Rentenversicherungsträgers
3-4	Geburtstag des Versicherten
5-6	Geburtsmonat des Versicherten
7-8	Geburtsjahr des Versicherten
9	Anfangsbuchstabe des Geburtsnamens des Versicherten
10-11	Seriennummer (männliche Versicherte = 00 bis 49, weibliche Versicherte = 50 bis 99)
12	Prüfziffer

Tabelle 3.2.: Aufbau der Rentenvericherungsnummer nach § 147 SBG VI [SGB08b]

Versichertenstatus

Pflichtmitglieder Zu den versicherungspflichtigen Personen zählen u.a. alle Beschäftigten, deren Einkommen die Versicherungspflichtgrenze nicht übersteigt, sowie Auszubildende, Praktikanten, Rentner, Künstler, Bezieher von Arbeitslosengeld I bzw. II, Übergangsgeld oder bestimmter anderer Entgeltersatzleistungen (vgl. § 5, SGB V [SGB08a]).

Familienmitversicherte Ehegatten bzw. eingetragene gleichgeschlechtliche Lebenspartner eines Versicherten, die über kein oder nur ein geringes Einkommen verfügen, können im Rahmen der gesetzlichen Krankenversicherung beitragsfrei mitversichert werden. Ebenso können Kinder bis zum 18. Lebensjahr über die Familienversicherung mitversichert werden. Junge Erwachsene, die das 23. Lebensjahr nicht vollendet haben und nicht erwerbstätig sind, werden ebenfalls familienmitversichert. Befindet sich das Kind in einer Schul- bzw. Berufsausbildung, wird die Familienversicherung bis zum 25. Lebensjahr verlängert (vgl. § 10, SGB V [SGB08a]).

Rentner Rentner sind nicht mehr erwerbstätige Personen, die ihren Lebensunterhalt aus einer Rente, also einer gesetzlichen oder privaten Versicherungsleistung, bestreiten.

freiwillig Versicherte Versicherte, die der Versicherungspflicht nicht unterliegen, werden als freiwillig Versicherte bezeichnet. Dazu zählen u.a. Beamte, Richter, Geistliche, Selbständige, Freiberufler, Soldaten und Arbeitnehmer, deren Einkommen die Versicherungspflichtgrenze übersteigt (vgl. § 6, SGB V [SGB08a]).

Struktur der Versichertenstammdaten-Tabelle

Die Versichertenstammdaten-Tabelle beinhaltet folgende Attribute:

Attribut	Bedeutung
KVNR	Krankenversicherungsnummer des Versicherten
RVNR	Rentenversicherungsnummer des Versicherten bzw. des Pflicht- bzw. freiwillig Versicherten (bei Familienmitversicherten für die noch keine Rentenversicherungsnummer generiert wurde, wird hier die RVNR des Hauptversicherten erfasst)
Angehörigennummer	0 bedeutet, es handelt sich um den "Hauptversicherten", ein Wert > 0 bedeutet, es handelt sich um einen Familienmitversicherten
Geburtsdatum	Geburtsdatum des Versicherten
Geschlecht	Geschlecht des Versicherten
PLZ	Postleitzahl des Wohnortes des Versicherten
Versichertenstatus	1: Pflichtversicherte oder freiwillig Versicherte 3: Familienmitversicherter 5: Rentner

KVNR	RVNR	Angehörigen-nummer	Geburtsdatum	Geschlecht	PLZ	Versichertenstatus
1234567	12110880T559	0	11.08.1980	w	46236	1
4563721	34290265G607	0	29.02.1965	w	41159	1
2453973	34290265G607	1	21.06.2000	m	58746	3
0253742	34290265G607	2	12.10.2003	w	59620	3
1672340	24290479J117	0	29.04.1979	m	78590	1
1237282	32301253H204	0	30.12.1953	m	86235	1

Abbildung 3.2.: Fiktive Beispieldatensätze der Versichertenstammdaten-Tabelle

3.4.2. Leistungsabrechnungsdaten

Wie bereits in Kapitel 3.3 beschrieben, rechnen alle Leistungserbringer die erbrachten Leistungen entweder direkt oder über entsprechende Verbände mit den Krankenkassen ab.

Abbildung 3.4.2 veranschaulicht die Datenflüsse von Versichertendaten im Rahmen der Leistungsabrechnung. Daten über erfolgte Leistungen (z.B. Hausbesuche, Blutentnahmen, Medikamente, gestellte Diagnosen) werden zunächst an die Kassen(zahn)ärztliche Vereinigung übermittelt. Diese werden dort gesammelt und, gemäß den Verträgen, aufbereitet. Anschließend werden diese an die Verbände der Krankenkassen übermittelt. Dort werden die Daten für die einzelnen Kassen aufbereitet und zur Verfügung gestellt (vgl. [Bol04]).

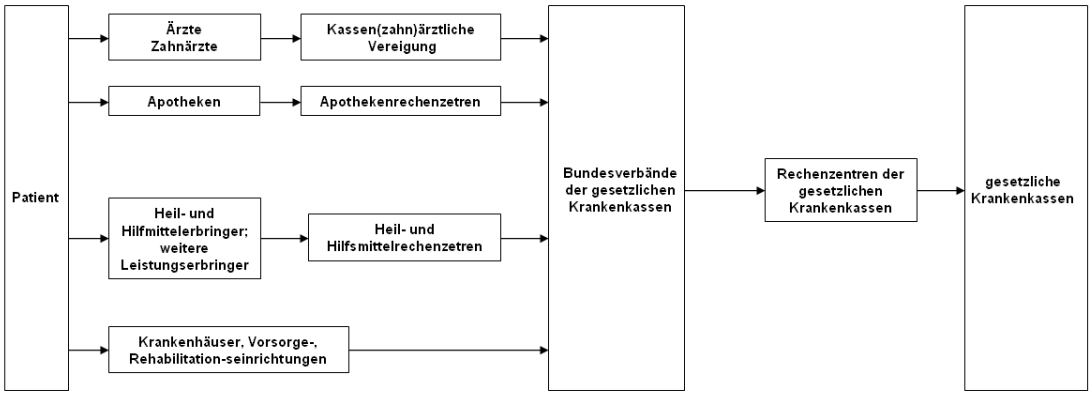


Abbildung 3.3.: Datenübermittlung [Gre06]

Internationale statische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme (ICD)

Die Internationale statische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme (ICD) ist eine von der Weltgesundheitsorganisation (WHO) herausgegebene internationale Diagnosenklassifikation von Krankheiten und verwandter Gesundheitsprobleme.

Die deutsche Übersetzung der ICD wird vom Deutschen Institut für Medizinische Dokumentation und Information (DIMDI) gepflegt und herausgegeben. Die aktuelle deutsche Ausgabe der ICD wird als ICD-10-GM bezeichnet.

Die ICD-10 gliedert sich hierarchisch in Krankheitskapitel, Krankheitsgruppen, Krankheitskategorien und Subkategorien. Jeder Diagnose oder Prozedur wird ein bis zu fünfstelliger Schlüssel zugeordnet. Aus den ersten drei Stellen einer ICD-Verschlüsselten Diagnose lässt sich die Krankheitskategorie ablesen, so werden im Rahmen dieser Untersuchungen nur dreistellige Diagnosen betrachtet. Diagnosen mit längerer Verschlüsselung werden auf drei Stellen gekürzt. Tabelle 3.3 gibt einen Überblick der ICD-10-Gliederung.

Nach § 295 und § 301 Sozialgesetzbuch V [SGB08a] sind Ärzte und Krankenhäuser verpflichtet, Diagnosen auf Abrechnungsunterlagen und Arbeitsunfähigkeitsbescheinigungen, sowie Krankenhausbehandlungen nach ICD zu verschlüsseln.

Arbeitsunfähigkeitsmeldung

Erkrankt ein Arbeitnehmer und kann infolgedessen seine Arbeit nicht ausüben, wird vom Arzt eine Arbeitsunfähigkeitsbescheinigung ausgestellt. Der Arbeitnehmer ist bei einer Erkrankung verpflichtet, die Arbeitsunfähigkeit unverzüglich beim Arbeitgeber und der Krankenkasse anzuzeigen und die voraussichtliche Dauer mitzuteilen. Die Krankenkasse benötigt diese Informationen, um den Anspruch und die Dauer des Krankengeldes zu berechnen.

ICD-10			Bedeutung
A00	-	B99	Bestimmte infektiöse Krankheiten
C00	-	D48	Neubildungen
D50	-	D90	Krankheiten des Blutes und der blutbildenden Organe sowie bestimmte Störungen mit Beteiligung des Immunsystems
E00	-	E90	Endokrine, Ernährungs- und Stoffelskrankheiten
F00	**-**	**F99**	**Psychische und Verhaltensstörungen**
G00	-	G99	Krankheiten des Nervensystems
H00	-	H59	Krankheiten des Auges und der Augenanhangsgebilde
H60	-	H95	Krankheiten des Ohres und des Warzenfortsatzes
I00	-	I99	Krankheiten des Kreislaufsystems
J00	-	J99	Krankheiten des Atmungssystems
K00	-	K93	Krankheiten des Verdauungssystems
L00	-	L99	Krankheiten der Haut und der Unterhaut
M00	-	M99	Krankheiten des Muskel-Skelett-Systems und des Bindegewebes
N00	-	N99	Krankheiten des Urogenitalsystems
O00	-	O99	Schwangerschaft, Geburt und Wochenbett
P00	-	P96	Bestimmte Zustände, die ihren Ursprung in der Perinatalperiode haben
Q00	-	Q99	Angeborene Fehlbildungen, Deformationen und Chromosomanomalien
R00	-	R99	Symptome und abnorme klinische und Laborbefunde, die anderenorts nicht klassifiziert sind
S00	-	T98	Verletzungen, Vergiftungen und bestimmte andere Folgen äußerer Ursachen
V01	-	Y98	Äußere Ursachen von Morbidität und Mortalität
Z00	-	Z99	Faktoren, die den Gesundheitszustand beeinflussen und zur Inanspruchsnahme des Gesundheitswesen führen
U00	-	U99	Schlüsselnummern für besondere Zwecke

Tabelle 3.3.: Hauptklassen der ICD-Klassifizierung nach [ICD06]

Struktur der Arbeitsunfähigkeitsmeldung-Tabelle Folgende Informationen stehen in der Arbeitsunfäfigkeitsmeldungs-Tabelle zur Verfügung:

Attribut	Bedeutung
RVNR	Rentenversicherungsnummer des Versicherten
KVNR	Krankenversicherungsnummer des Versicherten
VorgangsID	ID eines Vorgangs
Vorgang Von	Beginn der Arbeitsunfähigkeitsmeldung
Vorgang Bis	Ende der Arbeitsunfähigkeitsmeldung
ICD-10 Diagnose	ICD-10 verschlüsselte Diagnose
Diagnosetext	Klartext der ICD-10 Diagnose

KVNR	RVNR	VorgangsID	Vorgang Von	Vorgang Bis	ICD-10 Diagnose	Diagnosetext
1234567	12110880T559	1267	02.01.2002	12.01.2002	J35.8	Sonstige chronische Krankheiten der Gaumenmandeln und der Rachenmandel
4563721	34290265G607	2353	07.09.2003	07.09.2003	M62.6	Muskelzerrung
4563721	34290265G607	4875	01.11.2003	20.11.2003	J98.0	Krankheiten der Bronchien, anderenorts nicht klassifiziert
4563721	34290265G607	7562	25.03.2004	30.03.2004	N30.9	Zystitis, nicht näher bezeichnet
1672340	24290479J117	2362	03.03.2004	06.03.2004	K58.0	Reizdarmsyndrom mit Diarrhoe
1237282	32301253H204	2365	06.02.2005	18.02.2005	S63.5	Verstauchung und Zerrung des Handgelenkes

Abbildung 3.4.: Fiktive Beispieldatensätze der Arbeitsunfähigkeitsmeldung-Tabelle

Ambulantes Operieren

Unter dem Begriff "Ambulantes Operieren" wird eine operativen Eingriff verstanden, bei dem der Patient die Nacht vor und nach dem Eingriff zuhause verbringt. Mehr als 90 Eingriffe, wie z.B. Operationen des grünen oder grauen Stars, Knochenbruch-, Leistungsbruch-, Blinddarm-Operationen, Entfernung von Hauttumoren und Muttermalen, Herzschrittmacher- Implantationen, Entfernung von Mandeln, sowie Gebärmutterausschabungen (vgl. [AOP08]) können zur Zeit ambulant durchgeführt werden.

Die im Rahmen einer ambulanten Operation erbrachten Leistungen werden über die Krankenkasse abgerechnet.

Struktur der Ambulantes Operieren-Abrechnungsdaten-Tabelle Die Ambulantes Operieren-Abrechnungsdaten-Tabelle beinhaltet folgende Attribute:

Attribut	Bedeutung
KVNR	Krankenversicherungsnummer des Versicherten
RVNR	Rentenversicherungsnummer des Versicherten bzw. des "Hauptversicherten" bei Familienmitversicherten
Angehörigennummer	0 bedeutet, es handelt sich um den "Hauptversicherten", ein Wert > 0 bedeutet, es handelt sich um einen Familienmitversicherten
Vorgang Von	Beginn der Behandlung
Vorgang Bis	Ende der Behandlung
ICD 10-Schlüssel	ICD-10 verschlüsselte Diagnose

Krankenhausfälle

Krankenhäuser sind nach § 301 SGBV [SGB08a] gesetzlich verpflichtet, zu jeder Krankenhausbehandlung den Krankenkassen eine Reihe von Angaben ihrer Versicherten mitzuteilen, um zum einen eine ordnungsgemäße Abrechnung zu gewährleisten und zum

KVNR	RVNR	Angehörigen-nummer	Vorgang Von	Vorgang Bis	ICD-10 Diagnose
1234567	12110880T559	0	02.07.2003	02.07.2003	O02.1
4563721	34290265G607	0	12.03.2004	12.03.2004	N75.0
4563721	34290265G607	0	02.02.2005	02.02.2005	D48.6
4563721	34290265G607	1	25.06.2004	25.06.2004	Q52.1
1672340	24290479J117	0	29.01.2002	29.01.2002	D48.7
1237282	32301253H204	0	13.03.2005	13.03.2005	J35.2

Abbildung 3.5.: Fiktive Beispieldatensätze der ambulantes OperierenAbrechnungsdaten Tabelle

anderen die Notwendigkeit und die Dauer des Krankenhausaufenthalts zu überprüfen.

Struktur der Krankenhausfalldaten-Tabelle Folgende Informationen stehen in der Krankenhausfalldaten-Tabelle zur Verfügung:

Attribut	Bedeutung
KVNR	Krankenversicherungsnummer des Versicherten
RVNR	Rentenversicherungsnummer des Versicherten bzw. des "Hauptversicherten" bei Familienmitversicherten
Angehörigennummer	0 bedeutet, es handelt sich um den "Hauptversicherten", ein Wert > 0 bedeutet, es handelt sich um einen Familienmitversicherten
VorgangsID	ID eines Vorgangs
Vorgang Von	Beginn der Arbeitsunfähigkeitsmeldung
Vorgang Bis	Ende der Arbeitsunfähigkeitsmeldung
Fachabteilung	Station, die die Leistung erbracht hat
ICD 10-Schlüssel	ICD-10 verschlüsselte Diagnose
Diagnosetext	Klartext der ICD-10 Diagnose

KVNR	RVNR	Angehörigen-nummer	VorgangsID	Vorgang Von	Vorgang Bis	Fach-abteilung	ICD-10 Diagnose	Diagnosetext
1234567	12110880T559	0	1267	15.12.2003	17.12.2003	0100	I48	Vorhofflattern und Vorhofflimmern
4563721	34290265G607	0	2353	14.07.2003	15.07.2003	2200	N20.1	Ureterstein
4563721	34290265G607	0	4875	18.11.2004	18.11.2004	2800	R51	Kopfschmerz
4563721	34290265G607	1	7562	03.03.2005	07.03.2005	1000	A09	Diarrhoe und Gastroenteritis, vermutlich infektiösen Ursprungs
1672340	24290479J117	0	2362	25.11.2004	09.12.2004	0103	E78.0	Reine Hypercholesterinämie
1237282	32301253H204	0	2365	31.07.2003	06.08.2003	1500	K80.2	Gallenblasenstein ohne Cholezystitis

Abbildung 3.6.: Fiktive Beispieldatensätze der Krankenhausfalldaten-Tabelle

Rehabilitationsleistungen

Die medizinische Rehabilitation versucht, einen die Erwerbsfähigkeit bedrohenden oder entstandenen Gesundheitschaden (z.B. durch Unfall) zu beheben, zu mildern oder Folgen zu beseitigen. Medizinische Rehabilitation gibt es aber auch für Menschen, die nicht oder nicht mehr im Erwerbsleben stehen (z.B. Kinder oder alte Menschen) oder für Mütter und Väter (Mutter-/Vater-Kind-Kuren, Mütterkuren).

Die Gesetzliche Krankenversicherung finanziert Rehabilitationsleistungen, wenn diese erforderlich sind, um eine Krankheit zu erkennen, zu heilen, ihre Verschlimmerung zu verhüten oder Beschwerden zu lindern.

Struktur der Rehabilitationsleistungsabrechungsdaten-Tabelle Die Ambulantes Rehabilitationsleistungsabrechungsdaten-Tabelle beinhaltet folgende Attribute:

Attribut	Bedeutung
KVNR	Krankenversicherungsnummer des Versicherten
RVNR	Rentenversicherungsnummer des Versicherten bzw. des "Hauptversicherten" bei Familienmitversicherten
Angehörigennummer	0 bedeutet, es handelt sich um den "Hauptversicherten", ein Wert > 0 bedeutet, es handelt sich um einen Familienmitversicherten
Leistungszeitraum-Beginn	Beginn der Arbeitsunfähigkeitsmeldung
Leistungszeitraum-Ende	Ende der Arbeitsunfähigkeitsmeldung
Leistungserbringer	Bezeichnung des Leistungserbringers
ICD 10-Schlüssel	ICD 10 verschlüsselte Diagnose

KVNR	RVNR	Angehörigen-nummer	Vorgang Von	Vorgang Bis	ICD-10 Diagnose
1234567	12110880T559	0	23.03.2004	05.07.2004	J98.8
4563721	34290265G607	0	02.08.2005	23.08.2005	L20.8
4563721	35290350H107	0	15.09.2002	06.10.2002	A06.6
1672340	24290479J117	0	30.06.2004	21.07.2004	J06.9
1237282	32301253H204	0	18.11.2005	09.12.2005	M60

Abbildung 3.7.: Fiktive Beispieldatensätze der Kur-Abrechnungsdaten-Tabelle

3.4.3. Datenqualität

Im Folgenden wird die Qualität der einzelnen Daten erläutert.

Leistungsabrechnungsdaten

Insgesamt liegen 2.470.577 Leistungsabrechnungsdatensätze von 350.658 Versicherten einer Krankenkasse im Zeitraum von 2002 bis 2006 vor. Diese setzen sich zusammen aus:

- 1.494.684 Arbeitsunfähigkeitsmeldungen von 250.947 Versicherten,
- 28.644 Ambulantes Operieren-Abrechnungsdatensätzen von 23.163 Versicherten,
- 921.984 Krankenhausfällen von 122.662 Versicherten und
- 25.265 Rehabilitationsdatensätzen von 15.352 Versicherten

zusammen. 2.370 Datensätze konnten keinem Versicherten zugeordnet werden.

Versichertenstruktur

Das Versichertenverzeichnis beinhaltet Informationen von 943.031 Versicherten. Davon sind 706.574 pflichtversichert und 236.457 familienmitversichert.

Knapp $62,5\%$ der hier untersuchten Versicherten sind Frauen und das Durchschnittsalter beträgt $29,53$ Jahre. Abbildung 3.8 veranschaulicht die Geschlechter- und Altersverteilung.

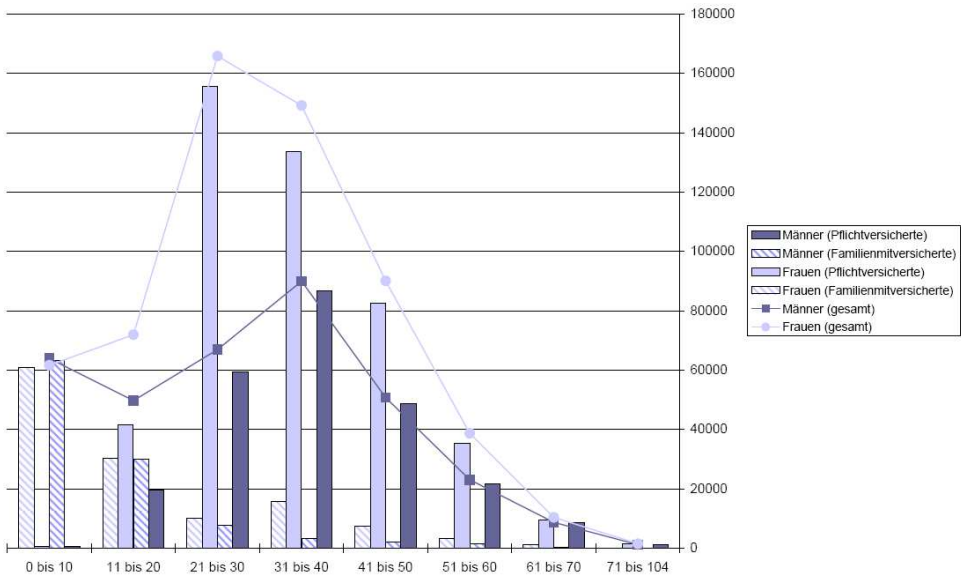


Abbildung 3.8.: Alters- und Geschlechterverteilung der Versicherten

Zu den am häufigsten ausgeübten Berufsgruppen der untersuchten Versicherten zählen Gesundheitsdienstberufe, wie z.B. Arzt und Sprechstundenhilfe, mit 41, 16%, Bürofach- und Bürohilfskräfte mit 9, 85% und Kaufleute mit 5, 70%.

Zu 592.373 Versicherten liegen keinerlei Abrechnungsinformationen vor, sodass keine Aussagen zu deren Krankheitsgeschichte getroffen werden können. Daher werden diese nicht weiter betrachtet. Von den übrigen 350.658 waren 287.591 Versicherte bezüglich psychischer Erkrankungen unauffällig.

Da ein Zeitfenster von 5 Jahren für die Analyse der Krankengeschichte von psychisch Kranken relativ klein ist, werden nur Versicherte betrachtet, die in den ersten 4 Jahren unauffällig bezüglich psychischer Erkrankungen waren und erst im letzten Jahr, also 2006, psychisch erkrankten. Dies trifft auf insgesamt 7.501 Versicherte zu.

Der Begriff "psychische Erkrankung" ist sehr weitläufig, er fasst alle Ausprägungen von psychischen Störungen, wie zum Beispiel organische psychische Störungen, Suchterkrankungen, Depressionen, Zwänge, Schizophrenie usw. zusammen. Die Krankengeschichte eines Versicherten, der an einer organischen psychischen Störungleidet, unterscheidet sich von der eines Suchtkranken. Aus diesem Grund ist es sinnvoll, Untergruppen von psychischen Störungen einzeln zu betrachten. Tabelle 3.4 listet die Untergruppen der Hauptklasse F der ICD-10-Klassifizierung auf.

ICD-10	Bedeutung
F00-F09	Organische, einschließlich symptomatischer psychischer Störungen
F10-F19	Psychische und Verhaltensstörungen durch psychotrope Substanzen, u.a. Suchtkrankheiten
F20-F29	Schizophrenie, schizotype und wahnhafte Störungen
F30-F39	Affektive Störungen, wie Manien oder Depressionen
F40-F48	Neurotische, Belastungs- und somatoforme Störungen, wie Phobien, Zwänge, Hysterie, Amnesie
F50-F59	Verhaltensauffälligkeiten mit körperlichen Störungen und Faktoren
F60-F69	Persönlichkeits- und Verhaltensstörungen
F70-F79	Intelligenzstörungen
F80-F89	Entwicklungsstörungen
F90-F98	Verhaltens- und emotionale Störungen mit Beginn in der Kindheit und Jugend
F99	Nicht näher bezeichnete psychische Störungen

Tabelle 3.4.: Untergruppen der ICD-10 F-Klasse

Abbildung 3.9 illustriert die Aufteilung der 7.501 Versicherten, die erst im letzten Jahr psychisch erkrankten, auf die ICD-10 F-Untergruppen.

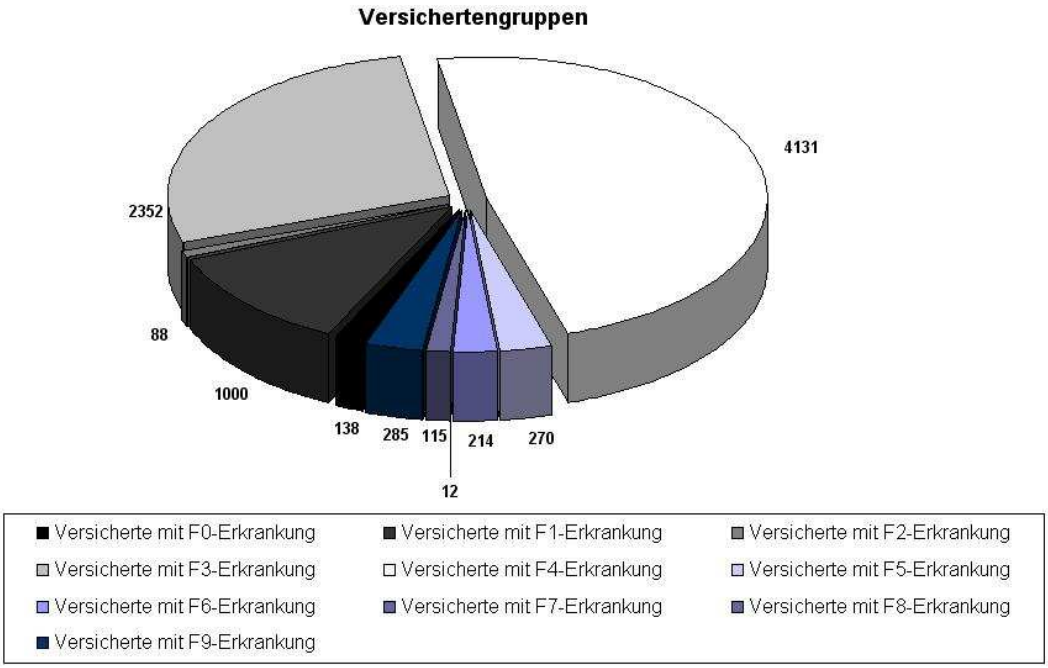


Abbildung 3.9.: Anzahl der Versicherten, die 2006 psychisch erkrankten, nach Untergruppen differenziert

Die vier größten Gruppen bilden:

- 4.131 Versicherte mit einer F4-Diagnose (Belastungs- und somatoforme Störungen),
- 2.352 Versicherte mit einer F3-Diagnose (affektive Störungen),
- 1.000 Versicherte mit einer F1-Diagnose (psychische und Verhaltensstörungen durch psychotrope Substanzen),
- 570 Versicherte mit einer F17-Diagnose (psychische und Verhaltensstörungen durch Tabak).

Potenzielle Fehlerquellen

An dieser Stelle werden mögliche Faktoren aufgezeigt, die die Ergebnisse negativ beeinflussen könnten.

kurzer Zeitraum Die vorliegenden Daten wurden im Zeitraum von 2002 bis 2006 gesammelt, somit kann nur ein relativ kleines Zeitfenster von 5 Jahren zur Analyse der Krankengeschichte herangezogen werden. Jedoch haben viele Störungen ihren Ursprung im des Kindes- und Jugendalter (vgl. [IELS04]).

keine lebenslanggültige Identifikationsnummer Zum Zeitpunkt des Datenabzugs wurde die ab Juli 2004 eingeführte lebenslanggültige Krankenversicherungsnummer (vgl. §290, SGB V [SGB08a]) noch nicht für alle Versicherten generiert, infolge-

dessen kann nicht gewährleistet werden, dass alle Daten einer Person korrekt zugeordnet werden können.

Als Beispiel sei ein 23 Jähriger Student genannt, der während seines Studiums nicht erwerbstätig ist. Er wird über einen Elternteil mitversichert und erhält eine Krankenversicherungsnummer. Eine Rentenversicherungsnummer wird aber nicht generiert, da kein Rentenversicherungskonto eröffnet wurde. Nach seinem Studium nimmt er eine Beschäftigung auf und es wird eine Rentenversicherungsnummer generiert. Gleichzeitig wird er zum Pflichtversicherten und eine neue Krankenversicherungsnummer wird erzeugt. In diesem Fall ist es nicht möglich alle Daten und folglich die Krankengeschichte des Studenten über den gesamten Zeitraum zusammenzuführen, da sowohl die Krankenversicherungsnummer als auch die Rentenversicherungsnummer neu generiert wurde.

fehlerhafte ICD-Kodierung Einer Umfrage zufolge rechnen noch knapp 7% der niedergelassenen Ärzte in Deutschland ihre erbrachten Leistungen bei den Krankenkassen über die Kassenärztliche Vereinigung in Papierform ab. Ungefähr ein Drittel der niedergelassenen Ärzte, die ein Praxis-Verwaltungs-Softwaresystem (PVS) für ihre Abrechnung nutzen, dokumentieren die gestellten Diagnosen zuerst in der Patientenkarte, bevor die Daten später von einer Arzthelferin in das PVS übertragen werden. Diverse Studien haben gezeigt, dass die Angaben, die direkt von dem behandelnden Arzt in die elektronische Patientenakte eingegeben wurden, vollständiger und korrekter waren als bei nachträglicher Eingabe von Notizen (vgl. [HHB+05] und [GPEP07]).

unvollständige Informationen zur Krankengeschichte Wie bereits in Kapitel 3.4.2 beschrieben, übermitteln die niedergelassenen Ärzte, um ihre Leistungen abzurechnen, ihre Daten zuerst an die für sie zuständige Kassenärztliche Vereinigung. Die Kassenärztliche Vereinigung wiederum muss die Daten aufbereiten, denn nach § 295 SGB V [SGB08a] dürfen für die Vergütungsabrechnung mit den Kassen keine Versichertennamen und Versichertennummern übermittelt werden. Die Daten werden nur fallbezogen und nicht versichertenbezogen abgerechnet. Das bedeutet, dass die Informationen über die in der Praxis gestellten Diagnosen und erbrachten Leistungen einem Versicherten nicht zugeordnet werden können. Dadurch kann lediglich auf Arbeitsunfähigkeitsmeldungen, Krankenhausfalldaten, Rehabilitations- und Ambulantes-Operieren-Abrechnunsdaten zurückgegriffen werden, da diese direkt an die Krankenkasse übermittelt werden.

4. Datenaufbereitung

Dieses Kapitel beschreibt die Selektion geeigneter Tabellen, Attribute und Datensätze, sowie die Ableitung höherwertiger oder andersartiger Attribute aus den ursprünglichen Attributen, die für die Krankengeschichte eines Versicherten relevant sind. Ebenso wird die notwendige Formattransformation für die in den nächsten Kapiteln beschriebenen Data-Mining-Algorithmen dargestellt.

4.1. Datenbereinigung

Zunächst wurden alle Tabellen von fehlerhaften Datensätzen bereinigt. Datensätze gelten als fehlerhaft, wenn

- eine Diagnose nicht der ICD-10-Codierung entspricht,
- der Datensatz keinem Versicherten im Versichertenverzeichnis zugeordnet werden kann oder
- kein bzw. ein ungültiger Leistungsabrechnungszeitraum vorliegt.

Insgesamt 6.620 Datensätze enthielten keine oder eine ungültige ICD-10-codierte Diagnose. 2.370 Leistungsabrechnungsdatensätze konnten keinem Versicherten zugeordnet werden und 1.572 Datensätze enthielten ein ungültiges Datum.

4.2. Datenselektion

Dieser Abschnitt beschreibt die Erzeugung neuer Attribute, die Auswahl relevanter Attribute, sowie die Generierung der Tabelle Diagnosen, die spätere Transformationen vereinfachen soll.

Abbildung 4.1 gibt vorab einen Überblick über alle durchgeführten Schritte der Datenaufbereitung. In den folgenden Kapiteln werden die einzelen Teilschritte näher erläutert.

4.2.1. Erzeugung neuer Attribute

Durch Transformation der ursprünglichen Attribute lassen sich neue Attribute generieren. Nachfolgend wird die Generierung von zwei neuen Attributen beschrieben.

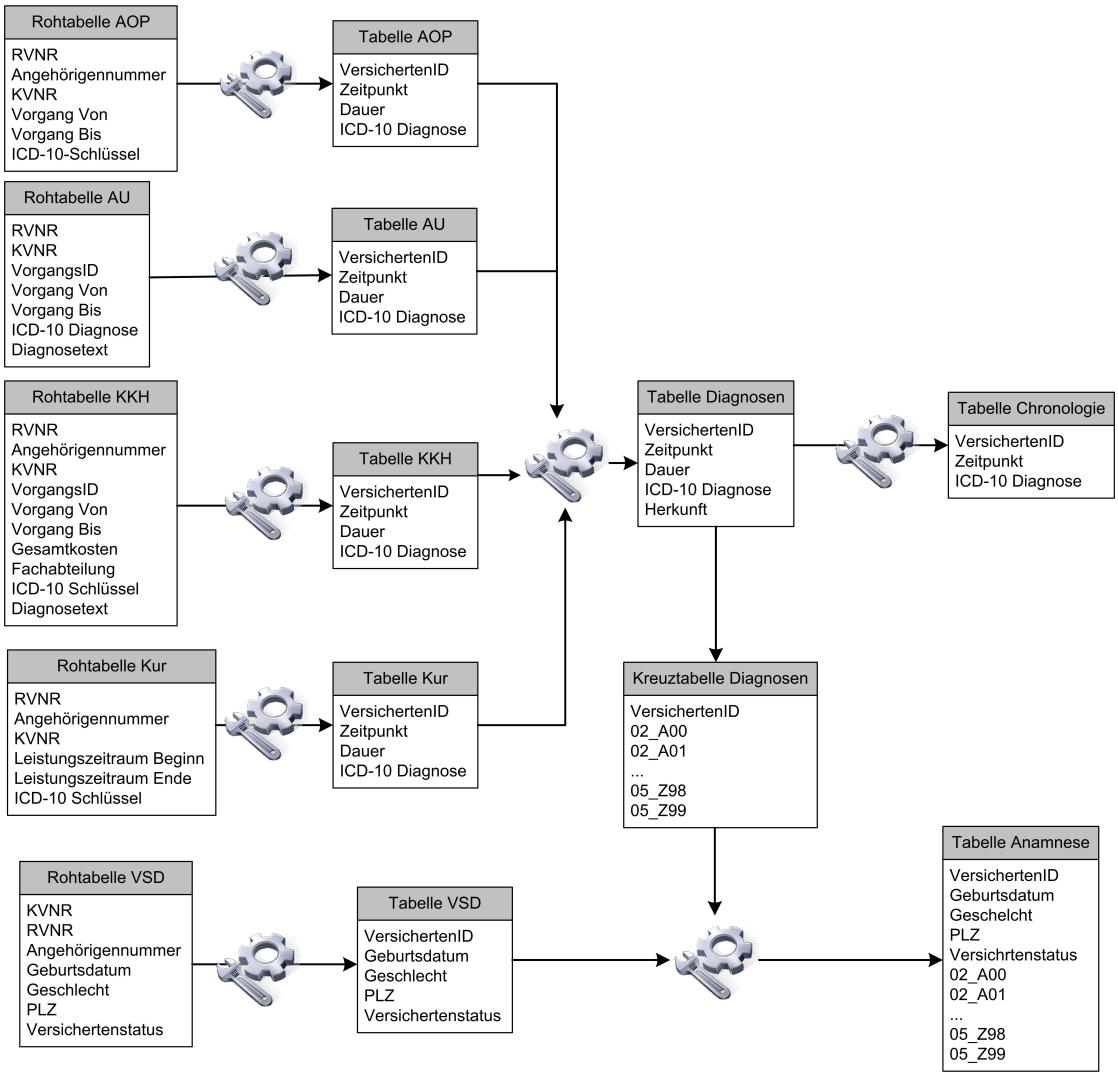


Abbildung 4.1.: Datenaufbereitung

Versicherten-Identifikationsnummer

Um die Daten eines Versicherten korrekt über einen möglichst langen Zeitraum zuordnen zu können, wird ein eindeutiges Attribut benötigt, das als Primärschlüssel verwendet werden kann.

Die Krankenversicherungsnummer ist nur bedingt dafür geeignet, da sie nicht lebenslang gültig ist und sich relativ häufig ändert.

Die Rentenversicherungsnummer ist zwar lebenslang gültig, aber nicht eindeutig, da Familienmitversicherte, für die keine eigene Rentenversicherungsnummer generiert wurde, unter derjenigen des Hauptversicherten (Ehegatten, Elternteil) geführt werden. Daher eignet sich die Kombination aus Rentenversicherungsnummer und Angehörigennummer

am besten als Primärschlüssel. Für Pflichtversicherte und Familienmitversicherte, für die bereits eine Rentenversicherungsnummer generiert wurde, wird an die zwölfstellige Rentenversicherungsnummer die Angehörigennummer 0 angehängt. Für alle anderen Familienmitversicherten wird die Angehörigennummer größer Null an die zwölfstellige Rentenversicherungsnummer des Hauptversicherten angehängt, siehe Abbildung 4.2.

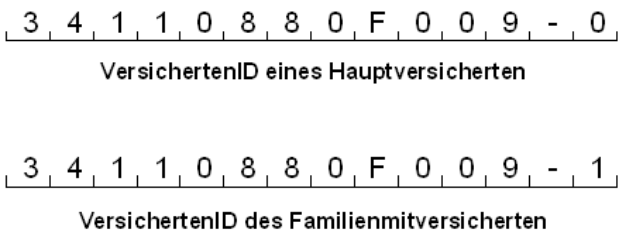


Abbildung 4.2.: Beispiel einer VersichertenID

Krankheitsdauer

Nicht nur der Zeitpunkt des Auftretens einer Krankheit ist für die Krankengeschichte relevant sondern auch die Dauer. Daraus lässt sich der Schweregrad der Krankheit ermitteln. In allen Tabellen ist der Beginn und das Ende eines Vorgangs erfasst, woraus sich die Krankheitsdauer in Tagen ermitteln lässt.

4.2.2. Auswahl geeigneter Attribute

Der Informationsumfang der Leistungsabrechungstabellen ist sehr unterschiedlich. Folgende Attribute bilden die Attributen-Schnittmenge aller Tabellen:

- VersichertenID
- Zeitpunkt
- Dauer
- ICD-10 codierte Diagnose

4.2.3. Struktur der Tabelle Diagnosen

Um spätere Transformationen zu vereinfachen, werden alle Datensätze in die Tabelle Diagnosen exportiert. Die neue Tabelle hat folgende Struktur:

Attribut	Bedeutung
VersichertenID	Versicherten-Identifikationsnummer
Zeitpunkt	Zeitpunkt des Auftretens der Erkrankung
Dauer	Dauer der Erkrankung
ICD-10 Diagnose	ICD-10 codierte Diagnose
Herkunft	Herkunft des Datensatzes AU: Arbeitsunfähigkeitsmelung, KKH: Krankenhaus Kur: Rehabilitationsabrechnungsdaten AOP: ambulantes Operieren-Abrechnungsdatensatz

4.3. Wissensrepräsentation

Als Anamnese wird die Darstellung der Vorgeschichte einer Erkrankung bezeichnet, die von einem Mediziner benötigt wird, um eine Diagnose stellen zu können. Eine Anamneseerhebung ist in vier große Bereiche untergliedert:

Krankheitsanamnese Unmittelbare Vorgeschichte der Erkrankung, insbesondere die Beschreibung der Symptomatik, der Beginn und eventuell der Auslöser der Erkrankung etc.

soziale Anamnese Detaillierte Abbildung der Lebensgeschichte, wie z.B. soziale Herkunft, berufliche Situation, Berufswahl und Ausbildung, Familienstand, Kinder, Schwangerschaften, längere Erkrankungen usw.

biografische Anamnese Chronologie der Erkrankungen

Familienanamnese Informationen über Familienangehörige, hinsichtlich Erbkrankheiten und Anfälligkeiten für bestimmte Erkrankungen

Für weiterführende Informationen siehe [Pay02].

4.3.1. Anamnesebasierte Darstellung

Informationen, die für einen Arzt zur Diagnostizierung einer Krankheit notwendig sind, können ebenfalls für die Beschreibung der Klassen der psychisch Kranken relevant sein. Aus den vorliegenden Versichertenstammdaten und den Leistungsabrechungsdaten können einige der benötigten Informationen gewonnen werden.

Die Versichertenstammdaten enthalten Informationen zum Versicherten selbst (Alter, Geschlecht, Wohnort, Familienangehörige) sowie zu seiner beruflichen Situation (Arbeitnehmer, Arbeitslos, Selbständig, vgl. Kapitel 3.4.1).

Informationen der Krankenanamnese sind in den Leistungsabrechungsdaten enthalten. Aus diesen Informationen lassen sich mehrere Arten von Darstellungen erzeugen. Im Folgenden werden fünf Darstellung beschrieben, die im Rahmen dieser Diplomarbeit untersucht wurden:

Vorerkrankungsübersicht Diese Darstellung gibt eine Übersicht über bisherige Vorerkrankungen eines Versicherten, ohne Angabe von Häufigkeit, Dauer und Zeitpunkt der Vorerkrankungen. Anhand dieser Darstellung können charakteristische Erkrankungen einer Versichertengruppen ermittelt werden.

Krankheitsdauersübersicht Die Krankheitsdauer ist ein Indikator für den Schweregrad einer Erkrankung. Eine lange Krankheitsdauer kann ebenso ein Hinweis auf Komplikationen sein. Diese Darstellung gibt einen Überblick darüber, wie lange ein Versicherter an einer bestimmten Krankheit behandelt wurde bzw. erkrankt ist.

Krankheitsdauer + Zeitangabe - Übersicht Diese Darstellung enthält die gleichen Informationen wie die Krankheitsdauerübersicht-Darstellung und zusätzlich die Information in welchem Jahr die Krankheit aufgetreten ist.

Häufigkeitsübersicht Eine hohe Häufigkeit des Auftretens einer Erkrankung kann ein Hinweis für ein schwaches Immunsystem oder schlechte Lebensbedingungen sein, aber auch Therapieabbrüche oder Konflikte mit Ärzten können zu erhöhter Häufigkeit einer Erkrankung führen. In [SE07] wird der Zusammenhang zwischen der erhöhten Häufigkeit und der Dauer von Erkankungen und somatoformen Störungen (ICD-10-Codierung: F45) hergestellt. Diese Darstellung gibt einen Überblick darüber, wie häufig ein Versicherter aufgrund einer bestimmten Diagnose behandelt wurde.

Häufigkeit + Zeitangabe-Übersicht Diese Darstellung enthält die identischen Informationen wie die Häufigkeitsübersicht, jedoch wird zusätzlich das Jahr des Auftretens der Krankheit mit angegeben.

Die Tabelle Diagnosen enthält alle Leistungsabrechungsdaten, jedoch wird nicht jede Diagnose in einem einzelnen Datensatz dargestellt. Anhand der "Häufigkeit + Zeitangabe"-Übersicht wird exemplarisch die Transformation der Leistungsabrechungsdaten in drei Schritten beschrieben. Die Generierung der anderen Darstellungen erfolgt analog.

Beispiel der Datentransformation

Um aus den Leistunsabrechnungsinformationen einen einzigen Datensatz pro Versicherten zu erzeugen, werden im ersten Schritt zwei neue Attribute erzeugt:

Attribut	Bedeutung
JJ_Diagnose	2-stellige Jahreszahl des Zeitpunktes des Auftretens der Diagnose kombiniert mit der 3-stellige ICD-10 Diagnose
Auftreten	konstant 1

Im zweiten Schritt werden alle Datensätze im Zeitraum von 1.1.2006 bis 31.12.2006 aus der Tabelle entfernt, da die Krankengeschichte der Versicherten nur im Zeitraum 2002 bis 2005 betrachtet werden soll.

Um eine Übersicht zu erhalten, die angibt, wie häufig innerhalb eines Jahres ein Versicherter an einer bestimmten Krankheit litt, muss im dritten und letzten Schritt das Auftreten der Krankheit summiert werden.

Auf diese Weise kann nun aus der erweiterten Tabelle Diagnosen die Kreuztabelle Diagnosen mit folgenden Attributen erzeugt werden:

Attribut	Bedeutung
VersichertenID	Versicherten-Identifikationsnummer
02_A00	$\sum_{\textbf{JJ\_Diagnose}=02\_A00}$ (**Auftreten**)
02_A01	$\sum_{\textbf{JJ\_Diagnose}=02\_A01}$ (**Auftreten**)
⋮	⋮
05_Z98	$\sum_{\textbf{JJ\_Diagnose}=05\_Z98}$ (**Auftreten**)
05_Z99	$\sum_{\textbf{JJ\_Diagnose}=05\_Z99}$ (**Auftreten**)

Abbildung 4.3 veranschaulicht die beschriebene Vorgehensweise.

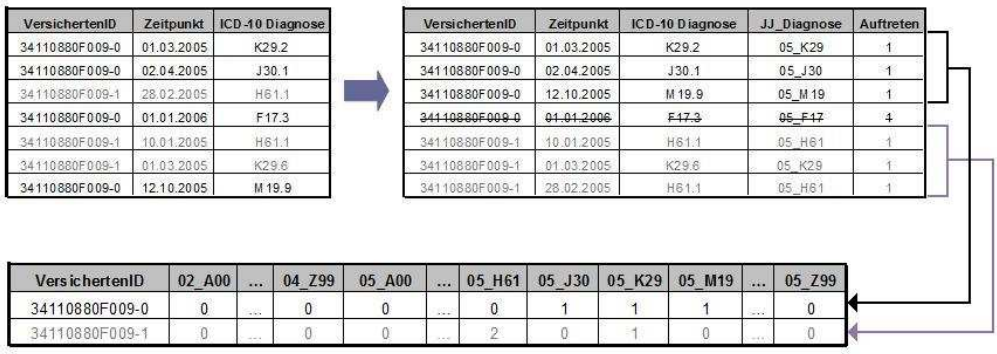


Abbildung 4.3.: Beispiel der Datentransformation bei der Krankheitsdauer-Darstellung

Um nun alle für die Anamnese benötigten Informationen (Informationen aus dem Versichertenverzeichnis und den Leistungsabrechnungsinformationen) zusammenzuführen, müssen die Tabelle VSD und die Kreuztabelle Diagnosen über die VersichertenID verbunden werden.

Tabelle4.2 stellt die auf diese Weise generierten Trainingsmengen dar.

4.3.2. Chronologische Darstellung der Vorerkrankungen

Bei dieser Darstellung der Krankengeschichte steht allein die Chronologie der Vorerkrankungen eines Versicherten im Vordergrund, um eventuell vorhandene Muster in der Krankengeschichte von psychisch Kranken erkennen zu können. So ein Muster könnte zum Beispiel Folgendes sein: zuerst tritt Krankheit x auf, dann tritt Krankheit y auf und dann wird mit 90% Wahrscheinlichkeit eine psychische Erkrankung z diagnostiziert. Stellt die Krankenkasse nun fest, dass ein Versicherter zuerst an Krankheit x und dann an Krankheit y erkrankt ist, könnte die Krankenkasse ihm ein speziell für die Krankheit z entwickeltes Präventionsprogramm anbieten.

Zur chronologischen Darstellung der Vorerkrankungen werden die Versichertenidentifikationsnummer, der Zeitpunkt des Auftretens der Erkrankung, sowie die ICD-10 codierte

Trainingsmenge	Anzahl Versicherte	Verhältnis F zu ohne F	Anzahl Attribute
F17-Diag. Vorerkrankungen	1.140	1 : 1	562
F17-Diag. Häufigkeit	1.140	1 : 1	562
F17-Diag. Häufigkeit + Zeit	1.140	1 : 1	1.091
F17-Diag. Krankheitsdauer	1.140	1 : 1	562
F17-Diag. Krankheitsdauer + Zeit	1.140	1 : 1	1.091
F1-Diag. Vorerkrankungen	2.000	1 : 1	672
F1-Diag. Häufigkeit	2.000	1 : 1	672
F1-Diag. Häufigkeit + Zeit	2.000	1 : 1	1.426
F1-Diag. Krankheitsdauer	2.000	1 : 1	672
F1-Diag. Krankheitsdauer + Zeit	2.000	1 : 1	1.426
F3-Diag. Vorerkrankungen	4.704	1 : 1	873
F3-Diag. Häufigkeit	4.704	1 : 1	873
F3-Diag. Häufigkeit + Zeit	4.704	1 : 1	873
F3-Diag. Krankheitsdauer	4.704	1 : 1	873
F3-Diag. Krankheitsdauer + Zeit	4.704	1 : 1	873
F4-Diag. Vorerkrankungen	8.262	1 : 1	1.025
F4-Diag. Häufigkeit	8.262	1 : 1	1.025
F4-Diag. Häufigkeit + Zeit	8.262	1 : 1	2.585
F4-Diag. Krankheitsdauer	8.262	1 : 1	1.025
F4-Diag. Krankheitsdauer+ Zeit	8.262	1 : 1	2.585
F-Diag. Vorerkrankungen	15.002	1 : 1	1.175
F-Diag. Häufigkeit	15.002	1 : 1	1.175
F-Diag. Häufigkeit + Zeit	15.002	1 : 1	3.194
F-Diag. Krankheitsdauer	15.002	1 : 1	1.175
F-Diag. Krankheitsdauer + Zeit	15.002	1 : 1	3.194

Tabelle 4.1.: Beschreibung der Trainingsmengen

Diagnose der Tabelle Diagnosen selektiert und nach VersichertenID und Zeitpunkt absteigend sortiert. Anschließend wird der Zeitpunkt des ersten Auftretens einer psychischen Erkrankung ermittelt und alle darauf folgenden Diagnosen des Versicherten entfernt, da nur die Vorgeschichte des Versicherten betrachtet werden soll. Abbildung 4.4 veranschaulicht die oben beschriebene Vorgehensweise.

Abbildung 4.4.: Beispiel der Datentransformation zur chronologischen Darstellung der Vorerkrankungen

Für jede Versichertengruppe (F1-Versicherte, F17-Versicherte, ..., Versicherte ohne psychische Erkrankung) werden auf diese Weise Testdaten für die später angewendeten Analysemethoden erzeugt. Tabelle 4.2 listet die daraus resultierenden Trainingsmengen auf.

Versicherten-gruppe	Anzahl Versicherte	Anzahl Diagnosen	Diagnosen pro Versicherten	max. Diagnosenanzahl eines Versicherten
ohne F-Diag.	7.501	48.812	6,508	162
mit F-Diag.	7.501	34.316	4,575	228
mit F1-Diag.	1.000	4.742	4,742	62
mit F17-Diag.	570	3.506	6,151	62
mit F3-Diag.	2.352	9.685	4,118	67
mit F4-Diag.	4.131	18.669	4,519	222

Tabelle 4.2.: Beschreibung der Trainingsmengen bei chronologischer Darstellung der Krankheiten

5. Lernverfahren

Dieses Kapitel beschäftigt sich zunächst mit dem Begriff des Lernens und geht anschließend auf die im Rahmen dieser Diplomarbeit eingesetzten Data Mining Verfahren ein.

5.1. Lernaufgaben

Gegenstand des maschinellen Lernens sind künstliche Systeme, die aus gegebenen Informationen lernen. Lernen bedeutet hierbei, dass sie in der Lage sind Lernaufgaben zu lösen. In [WMJ03] wird Lernaufgabe folgendermaßen definiert:

Definition 5.1 (Lernaufgabe) *Eine Lernaufgabe wird definiert durch:*

- *eine Beschreibung der dem lernenden System zur Verfügung stehenden Eingaben,*
- *der vom lernenden System erwarteten Ausgaben und*
- *den Randbedingungen des Lernsystems selbst.*

Ein System lernt erfolgreich genau dann, wenn es in der Lage ist, bei Eingaben, die den Spezifikationen entsprechen, unter den geforderten Randbedingungen Ausgaben mit den gewünschten Eigenschaften zu erzeugen.

Im Folgenden werden drei Lernaufgaben und Verfahren, welche diese lösen, vorgestellt.

5.2. Funktionslernen aus Beispielen

Die Aufgabe des Funktionslernens aus Beispielen besteht darin, aus einer Menge von Beispielen, welche durch einen Merkmalsvektor und ein Zielattribut, welches gemäß einer unbekannten Funktion berechnet wurde, beschrieben sind, eine Funktion zu lernen, welche die unbekannte Funktion möglichst gut approximiert, um für jedes neue Beispiel das Zielattribut möglichst gut berechnen zu können.

[WMJ03] definieren Funktionslernen aus Beispielen wie folgt:

Definition 5.2 (Funktionslernen aus Beispielen) *Es sei*

- *X eine Menge möglicher Instanzbeschreibungen,*
- *D eine Wahrscheinlichkeitsverteilung auf X,*
- *Y eine Menge möglicher Zielwerte,*
- *H eine Menge zulässiger Funktionen*

Eine Lernaufgabe vom Typ Funktionslernen aus Beispielen sieht dann wie folgt aus:

Gegeben: *Eine Trainingsmenge T von Beispielen, die gemäß der Wahrscheinlichkeitsverteilung D aus dem Instanzenraum X gezogen worden und mit einem Zielwert $y = f(x)$, einer unbekannten Funktion f versehen sind.*

Gesucht: *Eine Funktion $h \in H$, so dass der Fehler von h im Vergleich zu f minimiert wird.*

Um beurteilen zu können, wie gut die gelernte Funktion h die unbekannte Funktion f approximiert, muss ein geeignetes Fehlermaß herangezogen werden. Es gibt mehrere Alternativen die Güte der approximierten Funktion h zu messen.

Im Idealfall würde man den Fehler messen, welchen die Funktion h bei der Vorhersage der Zielwerte zukünftiger bzw. unbekannter Objekten macht. Dieses Gütekriterium wird $wahrerFehler$ genannt, der wie folgt definiert ist:

Definition 5.3 (Wahrer Fehler) *Sei D eine Wahrscheinlichkeitsverteilung auf dem Instanzenraum X, f die gesuchte Zielfunktion und $h \in H$. Als Wahren Fehler von h bezüglich f bezeichnet man:*

$$error_D(h) = Pr_{x \in X}[f(x) \neq g(x)]$$

Der wahre Fehler einer Funktion h ist also die Wahrscheinlichkeit, dass für ein beliebig bezüglich D gezogenes Beispiel x aus X der Zielwert $h(x)$ von $f(x)$ verschieden ist. Jedoch kann der wahre Fehler im Normalfall nicht berechnet werden, da weder die Funktion f noch die Verteilung D bekannt sind.

Alternativ zum wahren Fehler könnte der Fehler bestimmt werden, den die Funktion h bei der Bestimmung der Zielwerte der Trainingsmengebeispiele macht. Dieses Gütemaß wird $Trainingsfehler$ genannt und ist folgendermaßen definiert:

Definition 5.4 (Trainingsfehler) *Sei T eine Trainingsmenge von n Beispielen, f die gesuchte Zielfunktion und $h \in H$. Als Trainingsfehler von h bezüglich f bezeichnet man:*

$$error_T(h) := \frac{1}{n} \sum_{x \in T} error(f(x), h(x)),$$

$$mit\ error(f(x), h(x)) = \begin{cases} 0 & , falls\ h(x) = f(x) \\ 1 & , sonst. \end{cases}$$

Hierbei sind alle Kenngrößen bekannt und der Trainingfehler kann problemlos bestimmt werden. Allerdings tritt hier ein anderes Problem auf, nämlich das Problem der *Überanpassung*. Wird der Trainingsfehler beim Funktionslernen aus Beispielen als Minimierungskriterium verwendet, wird die Funktion h bestmöglich an die Trainingsmenge, welche nur eine (meist kleine) Teilmenge des Instanzraums repräsentiert, angepasst. Die so ermittelte Funktion h hat sich den Trainingsbeispielen sehr genau angepasst, erfasst jedoch nicht die unbekannte Funktion f.

Um dem Problem der Überanpassung vorzubeugen wird auf die $Kreuzvalidierung$ zurückgegriffen. Bei der Kreuzvalidierung wird die Trainingsmenge in m Teilmengen zerlegt. Eine Teilmenge zurückgehalten um das Ergebnis zu evaluieren, mit dem Rest wird gelernt. Dieses wird wiederholt bis alle Teilmengen einmal Evaluierungsmenge waren. Zur Schätzung des wahren Fehlers wird der Trainingsfehler über alle Teilmengen gemittelt. [WMJ03] definieren $m-fache-Kreuzvalidierung$ wie folgt:

Definition 5.5 (m-fache-Kreuzvalidierung) *Sei T eine Beispielsmenge, und L ein Lernverfahren. Dann partioniere E in m möglichst gleichgroße und disjunkte Teilmengen $T_1, \ldots, T_m$. Erzeuge nun die Hypothesen $h_1, \ldots, h_m$ wie folgt:*

$$h_i = \textit{ Ergebnis von } L \textit{ auf Basis der Beispielmengen } T\ T_i$$

Dann schätze den wahre Fehler von L auf T wie folgt:

$$error_{CV(T_1,\ldots,T_m)}(L,T) = \frac{\sum_{i=1,\ldots,m} error_{T_i}(h_i)}{m}$$

Umfassende Tests haben gezeigt, dass die 10-fach Kreuzvalidierung eine gute Schätzung für den wahren Fehler liefert.

Weitere Gütekriterien sind *Accuracy*, $Precision$, *Recall* und $F-Measure$ ([vR79]) diese werden jedoch nicht wie der Fehler minimiert, sonder maximiert.

Bevor die einzelnen Gütekriterien definiert werden, werden anhand einer zweidimensionalen Konfusionsmatrix alle vier möglichen Vorhersagen eines Klassifikationsproblems mit zwei Klassen vorgestellt.

		tatsächliche Klasse	
		positiv	negativ
ermittelte Klasse	positiv	TP	FP
	negativ	FN	TN

Tabelle 5.1.: 2D-Konfusionsmatrix für ein Klassifikationsproblem mit zwei Klassen

Dabei gibt TP die Anzahl der positiven Beispiele und TN die Anzahl der negativen Beispiele an, die richtig klassifiziert wurden. FP ist die Anzahl der Beispiele, die fälschlicherweise als positiv klassifiziert wurden und FN gibt die Anzahl der Objekte an, die fälschlicherweise negativ klassifiziert wurden.

Definition 5.6 (Accuracy) *Die Accuracy ist definiert als*

$$Accuracy := \frac{TP + TN}{TP + FP + FN + TN}.$$

Die *Accuracy* eines Lernverfahrens gibt die Wahrscheinlichkeit an, dass ein zufällig gezogenes Beispiel richtig klassifiziert wird.

Definition 5.7 (Recall) *Der Recall ist definiert als*

$$Recall := \frac{TP}{TP + FN}.$$

Unter *Recall* versteht man die Wahrscheinlichkeit, dass ein positives Beispiel auch als solches klassifiziert wird.

Definition 5.8 (Precision) *Die Precision ist definiert als*

$$Prescion := \frac{TP}{TP + FP}.$$

Precision ist die Wahrscheinlichkeit, dass ein positiv klassifiziertes Beispiel wirklich auch ein positives Beispiel ist.

Um verschiedene Experimente miteinander zu vergleichen, müssen $Precision$ und *Recall* gemeinsam betrachtet werden. Dazu wird ein Maß benötigt, welches *Recall* und $Precision$ kombiniert. Das $F - Measure$ kombiniert $Precision$ und *Recall*.

Definition 5.9 (F-Measure) *Das F-Measure ist definiert als*

$$F_\beta = \frac{(\beta^2 + 1) \cdot Precision \cdot Recall}{\beta^2 \cdot Precision + Recall}.$$

Der Parameter β bestimmt die Gewichtung von $Precision$ gegenüber *Recall*. Typischerweise wird $\beta = 1$ gesetzt um $Precision$ und *Recall* gleich zu gewichten.

Nachfolgend werden zwei Verfahren vorgestellt, welche die Lernaufgabe *Funktionslernen aus Beispielen* lösen.

5.2.1. Der Perzeptron-Algorithmus

Das Perzeptron wurde 1958 von Frank Rosenblatt [Ros58] entwickelt und später von Minsky und Papert [MP88] ausführlich untersucht. Das Perzeptron basiert auf der Idee eine biologische Nervenzelle mit einem mathematischen Modell nach zu bilden. Um die Funktionsweise eines Perzeptrons zu verstehen, ist es hilfreich, einen Blick auf das Vorbild zu werfen.

Das Neuron

Neuronen dienen im menschlichen Körper der Reizweiterleitung und -verarbeitung. Sie bestehen aus einem Perikaryon (Zellkörper), einem oder mehreren Dendriten (Zellfortsätzen) und einem Neurit, siehe Abbildung 5.1.

Dendriten sind feine, an das Perikaryon angrenzende Verästelungen. Sie dienen der Reizaufnahme durch angrenzende Synapsen anderer Nervenzellen. Die aufgenommen Reize werden in elektronische Impulse umgewandelt und an das Perikaryon weitergeleitet. Übersteigen die elektronischen Nervenimpulse einen bestimmten Schwellwert, werden diese über das Neurit an andere Zellen weitergeleitet, vgl. [BRDM97].

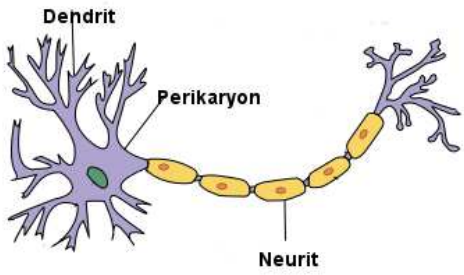
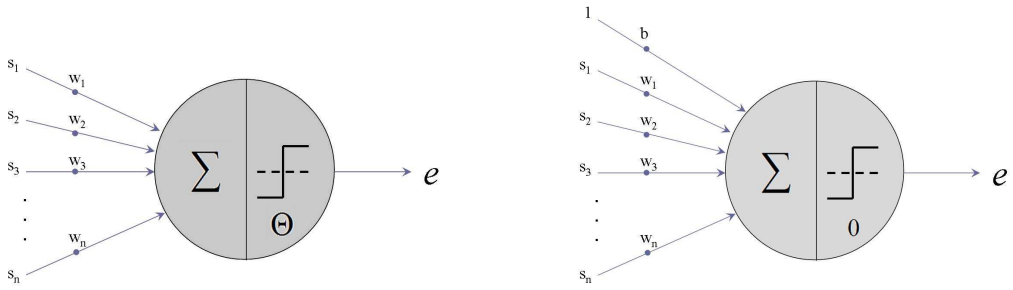


Abbildung 5.1.: Aufbau eines Neurons

Das Perzeptron

Das Perzeptron ist ein binäres Schaltelement mit gewichteten Eingängen, welches entweder aktiv oder inaktiv ist. Eingaben an den Eingängen des Perzeptrons entsprechen dabei den Eingangssignalen an den Dentriten des biologischen Neurons und der Funktionswert dem Ausgangsimpuls am Neurit.



(a) mit Schwelle Θ

(b) mit Schwelle 0 und $b = -\Theta$

Abbildung 5.2.: Einfaches Perzeptron

Die Verarbeitung der Eingangssignale zu einem Ausgangswert erfolgt dabei wie beim Vorbild in zwei Schritten:

Im ersten Schritt werden die Eingangssignale $s_1, \ldots, s_n$ anhand der Gewichtungen der Eingänge $w_1, \ldots, w_n$ verarbeitet, es wird die gewichtete Summe gebildet:

$$x = \sum_{i=1}^{n} w_i s_i = \mathbf{w} \cdot \mathbf{s} \tag{5.1}$$

Im zweiten Schritt wird überprüft, ob die Eingangssignale den Schwellwert überschreiten. Ist dies der Fall so wird das Perzeptron aktiv. Dazu wird die Aktivierungsfunktion auf die gewichtete Summe der Eingangswerte angewendet:

$$e := \begin{cases} 1 & \text{für } x \geq \Theta \\ 0 & \text{für } x < \Theta \end{cases} \tag{5.2}$$

mit $\mathbf{w}, \mathbf{s} \in \mathbb{R}^n$ und $\Theta \in \mathbb{R}$. Das n-dimensonale Perzeptron mit Schwellwert Θ lässt sich mit Hilfe eines Bias[1] $b = -\Theta$ in ein Perzeptron mit Schwellwert 0 überführen, siehe Abbildung 5.2 (b). Sei $\mathbf{w} = (b, w_1, \ldots, w_n)$ der Gewichtsvektor und $\mathbf{s} = (1, s_1, \ldots, s_n)$ der Signaleingansvektor, dann gilt:

$$e := \begin{cases} 1 & \text{für } \sum_{i=1}^{n} w_i s_i + b = \mathbf{w} \cdot \mathbf{s} + b \geq 0 \\ 0 & \text{für } \sum_{i=1}^{n} w_i s_i + b = \mathbf{w} \cdot \mathbf{s} + b < 0 \end{cases} \tag{5.3}$$

Die Gleichung $\mathbf{w} \cdot \mathbf{s} + b = 0$ definiert eine Hyperebene im Eingaberaum, mit dem Gewichtsvektor $\mathbf{w}$ als Normale der Hyperebenen. Somit ist das Perzeptron ein linearer Klassifikator, d.h es kann nur linear separierbare Funktionen berechnen, vgl. Abbildung 5.3.

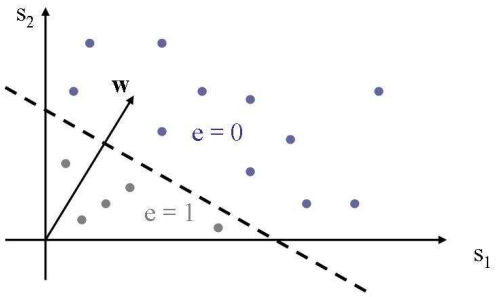


Abbildung 5.3.: Geometrische Interpretation des Perzeptrons

Der Lernalgorithmus

Gegeben sind zwei endliche Mengen S_1 und S_0. S_1 enthält alle $\mathbf{s}$ mit $f(\mathbf{s}) = 1$ und S_0 enthält alle $\mathbf{s}$ mit $f(\mathbf{s}) = 0$. Gesucht wird nun ein Gewichtsvektor $\mathbf{w}$ mit $\mathbf{s} \cdot \mathbf{w} \geq 0$ für alle $\mathbf{s} \in S_1$ und $\mathbf{s} \cdot \mathbf{w} < 0$ für alle $\mathbf{s} \in S_0$.

Der Gewichtsvektor $\mathbf{w}$ wird zunächst mit zufällig generierten Gewichten initialisiert. Anschließend werden alle Beispielsignale $\mathbf{s} \in S_1 \cup S_1$ iterativ ins Perceptron eingegeben, so lange bis das Perceptron alle Beispiele richtig klassifiziert. Jedes mal, wenn ein Eingangssignal falsch klassifiziert wird (das falsche Ausgangssignal erzeugt) muss der Gewichtsvektor w korrigiert werden. Die Gewichte werden dann nach folgender Regel angepasst:

$$w_i = w_i + n(t - o) \cdot x_i,$$

[1] *wikipedia:* Als Bias bezeichnet man in der Elektronik eine konstante, einseitige Größe, die meist absichtlich den eigentlichen Nutzsignalen überlagert wird.

wobei t der Sollwert und o der Istwert am Ausgangssignal des aktuellen Eingangssignals am Perceptron ist. n ist konstant und wird als Lernrate bezeichnet, diese legt die Stärke des Einflusses jeder Korrektur fest.

5.2.2. Naive Bayes

Der Naive Bayes-Klassifikator wurde erstmals 1973 von Duda und Hart ([DH73]) vorgestellt. Es handelt sich dabei um ist einen statistischen Klassifikator, der jedes Beispiel der Klasse zuordnet, zu der es mit der größten Wahrscheinlichkeit angehört. Der Naive Bayes-Klassifikator basiert auf dem Bayes'schen Theorem und der naiven Annahme, dass alle Attribute stochastisch unabhängig von einander sind.

Bayes' Theorem

Das Bayes' Theorem ist nach Thomas Bayes, einem bekannten Theologen und Mathematiker des 18.Jahrhunderts, benannt, der u.a. durch folgenden Satz bekannt wurde:

Satz 5.1 (Satz von Bayes) *Die Ereignisse $A_1, \ldots, A_n$ seien paarweise disjunkt. Ferner sei $B \subseteq A_1 \cup \ldots A_n$ ein Ereignis mit $Pr(B) > 0$. Die Wahrscheinlichkeit, dass eines dieser Ereignisse unter der Bedingung von B eintritt, ist*

$$Pr(A_k|B) = \frac{Pr(B|A_k) \cdot Pr(A_k)}{Pr(B)}$$

Dabei ist $Pr(A_k)$ die Apriori-Wahrscheinlichkeit für das Ereignis A_k, $Pr(B|A_k)$ die Wahrscheinlichkeit für das Ereignis B unter der Bedingung, dass A eingetreten ist und $P(B)$ die Apriori-Wahrscheinlichkeit für das Ereignis B.

Klassifikation mit Naive Bayes

Gegeben ist eine Trainingsmenge T von Beispielen, welche durch m Merkmale beschrieben und einer von n Klassen zugeordnet sind. Ein noch nicht klassifiziertes Beispiel, welches durch den Merkmalsvektor $\mathbf{x} = (x1, \ldots, x_m)$ beschrieben ist, soll nun einer der n Klassen zugeordnet werden. Dazu wird zunächst für jede der n Klassen die Wahrscheinlichkeit berechnet, dass das zu klassifizierende Beispiel mit dem Merkmalsvektor $\mathbf{x}$ der Klasse C_i zugeordnet wird, also $Pr(C_i|\mathbf{x})$. Das Beispiel wird der Klasse zugeordnet für die $Pr(C_i|\mathbf{x})$ maximal ist.

$$h(\mathbf{x}) = argmax_{C_i \in C} Pr(C_i|\mathbf{x}) \tag{5.4}$$

Um $Pr(C_i|\mathbf{x})$ zu berechnen wird der Satz von Bayes eingesetzt, somit gilt:

$$Pr(C_i|\mathbf{x}) = \frac{Pr(\mathbf{x}|C_i) \cdot Pr(C_i)}{Pr(\mathbf{x})} \tag{5.5}$$

$Pr(\mathbf{x})$ ist die Apriori-Wahrscheinlichkeit, dass ein Beispiel durch den Merksmalvektor $\mathbf{x}$ beschrieben wird. Diese ist für allen Klassen konstant, also muss nur $Pr(\mathbf{x}|C_i) \cdot Pr(C_i)$ maximiert werden.

$Pr(C_i)$ ist die Apriori-Wahrscheinlichkeit, dass ein Beispiel der Klasse C_i zugeordnet wird. Die Apriori-Wahrscheinlichkeit der einzelnen Klassen kann folgendermaßen aus der Trainingsmenge geschätzt werden:

$$Pr(C_{i,T}) = \frac{|C_{i,T}|}{|T|} \tag{5.6}$$

Dabei ist $|C_{i,T}|$ die Anzahl der Trainingsbeispiele, welche der Klasse C_i zugeordnet sind, und $|T|$ die Gesamtzahl der Trainingsbeispiele.

Nun muss noch die Wahrscheinlichkeit $Pr(\mathbf{x}|C_i)$ bestimmt werden. Um dies zu vereinfachen, wird die "naive" Annahme gemacht, dass alle Attribute stochastisch unabhängig sind, d.h. sich nicht gegenseitig beeinflussen, daher auch der Name Naive Bayes. Unter dieser Annahme kann $Pr(x|C_i)$ nun einfach aus der Trainingsmenge berechnet werden:

$$Pr(\mathbf{x}|C_i) = \prod_{k=1}^{n} Pr(x_k|C_i) \tag{5.7}$$

5.3. Subgruppenentdeckung

Ziel der vorangegangenen Lernaufgabe war es ein globales Modell zu finden, welches für jede mögliche Instanz eine möglichst gute Vorhersage machen kann. Im Gegensatz dazu besteht die Aufgabe der Subgruppenentdeckung darin, interessante lokale Modelle zu finden, die Teilbereiche der Trainingsmenge beschreiben. Für die Aufgabenstellung dieser Diplomarbeit wäre folgende fiktive Aussage, die eine Versichertengruppe beschreibt, interessant: "Unter den familienmitversicherten Jungen, im Alter von 5 bis 15 Jahren ist die Wahrscheinlichkeit signifikant höher als im gesamten Versichertenstamm an einer einfachen Aktivitäts- und Aufmerksamkeitsstörung (F90) zu leiden." Abbildung 5.4 soll die unterschiedliche Verteilung des Zielattributes "F90" in der Gesamtpopulation (gesamte Versichertenstamm) und der Subgruppe "familienmitversicherte Jungen im Alter von 5 bis 15 Jahren" verdeutlichen.

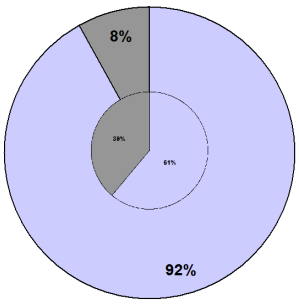


Abbildung 5.4.: Verteilung des Zielattributes in der Gesamtpulation und der Subgruppe

[WMJ03] definieren Subgruppenentdeckung wie folgt:

Definition 5.10 (Subgruppenentdeckung) *Sei X ein Instanzraum mit einer Wahrscheinlichkeitsverteilung D und L_H ein Hypothesenraum, in dem jede Hypothese als Extension eine Teilmenge von X hat: $ext(h) \subset X$ für alle $h \in L_H$. Sei weiterhin $S \subset X$ eine gegebene, gemäß D gezogene Stichprobe der Gesamtpopulation. Es sei schließlich q eine Funktion $q := L_H \rightarrow \mathbb{R}$.*

Die Lernaufgabe Subgruppenentdeckung kann dann auf zwei Arten definiert werden:

1. *Gegeben X, S, L_H, q und eine Zahl $q_{min} \in \mathbb{R}$, finde alle $h \in L_H$, für die $q(h) \geq q_{min}$. Und/Oder*
2. *gegeben X, S, L_H, q und eine natürliche Zahl $k \geqq 1$, finde eine Menge $H \subset L_h$, $|H| = k$ und es gibt keine $h \in H$, $h' \in L_H$ $H : q(h') \geq q(h)$.*

Die Aufgabe der Subgruppenentdeckung besteht also darin alle lokalen Modelle zu identifizieren, deren Qualität den Schwellenwert q_{min} übersteigt und/oder die k bezüglich q besten lokalen Modelle zu finden.

Eine wichtige Rolle bei der Subgruppennentdeckung ist die Auswahl einer geeigneten Qualitätsfunktion q.

Nachfolgend werden zwei Funktionen vorgestellt, welche die Interessantheit einer Subgruppe bewerten.

Definition 5.11 (Bias) *Sei $A \rightarrow B$ eine Regel[2], welche eine Subgruppe beschreibt. Dann ist der Bias der Regel definiert als*

$$BIAS(A \rightarrow B) := Pr[B|A] - Pr[B]$$

Der *Bias* ist somit ein Maß für die Abweichung der Verteilung des Zielattributes in der Subgruppe im Vergleich zu der Gesamtpopulation.

Lift ist das inverse Gegenstück zu *Bias* und ist wie folgt definiert:

Definition 5.12 (Lift) *Sei $A \rightarrow B$ eine Regel, welche eine Subgruppe beschreibt. Dann ist der Lift der Regel definiert als*

$$LIFT(A \rightarrow B) := \frac{Pr[B \cap B]}{Pr[A] \cdot Pr[B]} = \frac{Pr[B|A]}{Pr[B]}$$

5.3.1. Knowledge-Based Sampling

Das von [Sch05] vorgestellte Verfahren *Knowledge-Based Sampling* löst das Problem der Entdeckung interessanter Subgruppen. Das iterative Verfahren basiert auf der Idee bereits bekanntes Wissen aus dem Gesamtpopulation zu entfernen, um neue interessante

[2] Häufig werden Hornklauseln [Stö95]zur Beschreibung von Subgruppen verwendet.

Subgruppen zu entdecken. Dies geschieht indem die Wahrscheinlichkeit ein Beispiel zu ziehen verändert wird.

Zu Beginn, wenn noch kein Vorwissen bekannt ist, werden alle Beispiele mit gleicher Wahrscheinlichkeit gezogen und alle Beispiele sind gleichgewichtet. Beim ersten Durchlauf wird nun ein signifikanter Unterschied der Verteilung des Zielattributes in einer Subgruppe im Vergleich zur Gesamtpopulation entdeckt. Dies bedeutet, dass zwischen dieser Subgruppe und dem Zielattribut eine Korrelation besteht. Um nun weitere interessante Subgruppen zu entdecken, muss die Verteilung neu konstruiert werden, so dass die Korrelation zwischen der zuerst entdeckten Subgruppe und dem Zielattribut nicht mehr existent ist.

Folgende Bedingungen sind dabei für die neue Verteilung D' besonders zu beachten:

- Die Verteilung des Zielattributes in der Subgruppe, soll der Verteilung des Zielattributes in der Gesamtpopulation angepasst werden:
  $Pr_{D'} = (B|A) = Pr_{D'}(B)$.
- Die Wahrscheinlichkeit ein Beispiel aus der Subgruppe zu ziehen, soll unverändert bleiben:
  $Pr_{D'}(A) = Pr_D(A)$.
- Die Wahrscheinlichkeit ein Beispiel mit dem Zielattribut zu ziehen, soll gleich bleiben: $Pr_{D'}(B) = Pr_D(B)$.

Weitere Anforderungen sind [Sch05] zu entnehmen.

Um die neue Verteilung zu erhalten, werden die Gewichtungen der einzelnen Beispiele angepasst. Dazu wird zunächst der $LIFT$ der einzelnen Beispiele $x \in X$ ermittelt:

$$LIFT(x, A \to B) := \begin{cases} LIFT(A \to B) & \text{, falls } x \in A \cap \bar{B} \\ LIFT(A \to \bar{B}) & \text{, falls } x \in A \cap \bar{B} \\ LIFT(\bar{A} \to B) & \text{, falls } x \in \bar{A} \cap B \\ LIFT(\bar{A} \to \bar{B}) & \text{, falls } x \in \bar{A} \cap B \end{cases}.$$

Die neue Gewichtung wird nun wie folgt berechnet:

$$w_{t+1}(x) = \frac{w_t(x)}{LIFT(x, A \to B)}$$

Damit ergibt sich die neue Verteilung D' folgendermaßen:

$$Pr_{D'}(x) = \frac{Pr_D(x)}{LIFT(x, A \to B)}$$

Der Nachweis der Richtigkeit der Formel wird in [Sch05] erbracht.

Abbildung 5.3.1 veranschaulicht die gerade beschriebene Vorgehensweise in Pseudocode-Notation.

Algorithm 1 KBS

Require: Beispielmenge $E = \langle x_1, y_1 \rangle, \ldots, \langle x_1, y_1 \rangle$ mit $x \in X$ und $y =$ Zielwert, Anzahl der zu suchenden lokalen Modelle n

D_1 bezeichnet die Gleichverteilung über E.

for $i = 1, \ldots, n$ **do**

$h_i :=$ Lerner(D_i, E)

Berechne $LIFT_{D_i}(x_i, h_i)$

$D_{i+1} := D_i(x_i) \cdot (LIFT_{D_i}(x_i, h_i))^{-1}$

end for

Zur Durchführung der Experimente wurde der BayesianBoosting RapidMiner-Operator verwendet. Dieser liefert mit folgender Parameterbelegung *rescale label priors = True , allow marginal skews = False* die gleichen Ergebnisse wie der oben beschriebene KBS-Algorithmus. Der Parameter *iterations = 10* begrenzt die Anzahl der zu suchenden lokalen Modelle auf 10. Als interne Basislerner wurden zum einen das in Abschnitt 5.2.1 beschriebene Perceptron und zum anderen der DecisionTree (vgl. [BFOS84] und [BFOS84]) eingesetzt.

5.4. Entdeckung häufiger Sequenzen

Eine weitere vielfach untersuchte Lernaufgabe ist die Suche nach häufigen Sequenzen in Datenbanken, dabei werden in großen Datenbanken zeitliche Ereignismuster gesucht.

Seit der großflächigen Einführung von Barcodes in Supermärkten Ende der 70er Jahre, stehen den Betreibern riesige Mengen von Verkaufsdaten zur Verfügung. Solche Verkaufsdaten enthalten typischerweise Informationen über den Zeitpunkt des Einkaufs sowie über die Waren, welche gekauft wurden. Verwendet der Kunde eine Kreditkarte oder eine so genannte Kundenkarte, können diese Verkaufsdaten sogar einem Kunden zugeordnet werden. Die auf diese Weise personalisierten Verkaufsdaten spiegeln das Kaufverhalten eines Kunden wider, siehe Tabelle 5.2.

Kunde	Zeitpunkt	gekaufte Artikel (*Items*)
Kunde 1	01.04.05	Grill, Holzkohle
Kunde 1	08.04.05	Chips, Gummibärchen
Kunde 1	10.04.05	Würstchen, Bier, Ketchup
Kunde 2	01.04.05	Grill, Grillzange
Kunde 2	02.04.05	Kartoffel, Eier
Kunde 2	05.04.05	Würstchen, Ketchup, Limonade

Tabelle 5.2.: Kundendaten

Die Betreiber von Supermärkten sind sehr an der Analyse solcher Verkaufsdaten interessiert, um für sie aufschlussreiche Zusammenhänge zu entdecken. Dem obigen Beispiel ist folgender Zusammenhang zu entnehmen: *Kunden die einen Grill kaufen, werden im*

weiteren Verlauf Würstchen und Ketchup kaufen. Durch Entdeckung solcher zeitlicher Muster im Kaufverhalten der Kunden, können sie ihre Geschäftsprozesse optimieren. Vor diesem Hintergrund untersuchten Agrawal und Srikant 1995 das Problem der Entdeckung häufiger Sequenzen in Datenbanken (vgl.[AS95] und [SA96]).

Um die Lernaufgabe *Entdeckung häufiger Sequenzen* sowie ein Verfahren, der diese Lernaufgabe löst, zu beschrieben werden, bedarf es zunächst folgender Definitionen.

Der Einkauf eines Kunden entspricht einer *Transaktion.* Eine Transaktion ist ein Tripel $T = (ID_K, t, I)$ mit $ID_k, t \in \mathbb{N}$ und I eine nicht-leere Menge von Items. Dabei ist ID_K die eindeutige Identifikationsnummer eines Kunden, t der Zeitpunkt des Einkaufs und I gekauften Artikels. Das Tupel $(ID_K), t)$ bildet den Identifikationsschlüssel einer Transaktion. Eine nicht-leere Menge von Transaktionen wird als *Transaktionsdatenbank* bezeichnet.

Definition 5.13 *Ein Itemset ist eine nicht-leere Menge $I := (i_1, i_2, \ldots, i_m)$ von Items i_j. Eine Sequenz S ist eine geordnete Liste von Itemsets $\langle I_1 I_2 \ldots I_n \rangle$, wobei I_j ein Itemset darstellt.*

Werden alle Transaktionen eines Kunden zusammengefasst und nach der Transaktionszeit geordnet, bilden sie eine Sequenz. Diese Sequenz wird Kundensequenz genannt.

Definition 5.14 *Eine Sequenz $A = \langle A_1 \ldots A_n \rangle$ ist in der Sequenz $B = \langle B_1 \ldots B_m \rangle$ enthalten, falls es ganze Zahlen der Form $i_1 < i_2 < \ldots < i_n$ gibt, so dass gilt:*

$$A_1 \subseteq B_{i_1}, A_2 \subseteq B_{i_2}, \ldots, A_n \subseteq B_{n_1},$$

d.h. jedes Itemset von A ist in einem Itemset von B enthalten.

Ein Beispiel soll die Definition verdeutlichen: $A = \langle (3)(4\ 5)(8) \rangle$ ist in $B = \langle (7)(3\ 8)(9)(4\ 5\ 6)(8) \rangle$ enthalten, da $(3) \subseteq (3\ 8)$, $(4\ 5) \subseteq (4\ 5\ 6)$ und $(8) \subseteq (8)$. Im Gegensatz dazu ist die Sequenz $C = \langle (3)(5) \rangle$ nicht in $D = \langle (3\ 5) \rangle$ enthalten.

Definition 5.15 *Der Support $sup(S)$ einer Sequenz S ist definiert als die Anzahl der Kundensequenzen, die S enthalten. In Bezug auf eine feste Zahl s_{min} ist eine Sequenz häufig genau dann, wenn $sup(S) \geq s_{min}$ gilt.*

Das Problem der *Entdeckung von häufigen Sequenzen* lässt sich nun wie folgt definieren:

Definition 5.16 (Entdeckung von häufigen Sequenzen) *Es sei*

- *D eine nicht-leere Menge von Transaktionen,*
- *$s_{min} \in \mathbb{N}$ eine benutzergegebene Minimalhäufigkeit (minimum support),*

Eine Lernaufgabe vom Typ Entdeckung von häufigen Sequenzen sieht dann wie folgt aus:

Gegeben: *D, s_{min}*

Gesucht: *Menge $M_{s_{min}}$ der häufigen Kundensequenzen $S \in D$, für die gilt:*
$M_{s_{min}} = \{S| \geq s_{min} sup(S) s_{min}\}$.

5.4.1. Der GSP-Algorithmus

Der von Agrawal und Srikant 1996 entwickelte GSP-Algorithmus [SA96], welcher die oben beschriebene Lernaufgabe löst, basiert auf dem Prinzip der Kanidaten-Generierung.

Der GSP-Algorithmus durchläuft mehrere Iterationen. Bei der ersten Iteration wird die gesamte Transaktionsmenge D durchlaufen und für jedes Item wird eine Sequenz der Länge 1 gebildet, gleichzeitig wird der Support $sup(S)$ der 1-Sequenzen bestimmt. Anschließend werden die Sequenzen, welche die Minimalhäufigkeit s_{min} nicht erfüllen, aus der Grundmenge für den nächsten Durchlauf entfernt. Im nächsten Durchlauf werden aus dieser Grundmenge die Kandidaten für häufige Sequenzen der Länge 2 generiert. Anschließend wird der Support der Kandidaten ermittelt. Am Ende werden die Sequenzen, die den s_{min} nicht erfüllen aus der Grundmenge für den nächsten Durchlauf entfernt. Diese drei Phasen (Kandidatengenerierung, Bestimmung des Supports, Entfernung der Sequenzen, die den s_{min} erfüllen) wird so lange wiederholt bis keine neuen Kandidaten für häufige Sequenzen generiert werden können.

Abbildung 5.4.1 veranschaulicht die gerade beschriebene Vorgehensweise in Pseudocode-Notation.

Algorithm 2 GSP-Algorithmus

Require: Transaktionsmenge D, Minimalhäufigkeit s_{min}
  C_1 := GenerateIntialCanidates(D)
  C_1 := Prune(C_1)
  $k := 1$
  while $C_k \neq \emptyset$ **do**
    C_{k+1}:=GenerateCanidates(C_k)
    CountSupport($C_k + 1$, D)
    $k := k + 1$
  end while

GenerateCanidates

In dieser Phase werden aus Sequenzen der Länge k Kandidaten der Länge $k+1$ erzeugt, dies geschieht in folgenden zwei Teilschritten:

Join Zunächst zwei Sequenzen $S = s_1, s_2, \ldots, s_k$ und $T = t_1, t_2, \ldots, t_k$ der Länge k werden zu einer Sequenz der Länge $k+1$ verbunden, falls $s_2 = t_1, s_3 = t_2, \ldots, s_k = t_{k-1}$ gilt. Die neu erzeugte Kandidatensequenz der Länge $k+1$ entsteht indem der Sequenz S das letzte Item der Sequenz T hinzugefügt wird.

Prune In diesem Schritt werden die Sequenzen $c \in C_{k+1}$ entfernt, deren Teilsequenzen der Länge k nicht in C_k enthalten sind.

Tabelle 5.3 veranschaulicht die beschrieben Vorgehensweise anhand eines Beispiels.

Auf den Beweis der Richtigkeit dieser Prozedur wird an dieser Stelle verzichtet und auf [AS95] verwiesen.

häufige 3-Sequenzen	Kandidaten der Länge 4 (nach Join)	Kandidaten der Länge 4 (nach Prune)
⟨1 2 3⟩	⟨1 2 3 4⟩	⟨1 2 3 4⟩
⟨1 2 4⟩	⟨1 2 4 5⟩	
⟨1 3 4⟩	⟨1 3 4 5⟩	
⟨1 3 5⟩	⟨1 3 5 4⟩	
⟨2 3 4⟩		

Tabelle 5.3.: GenerateCanidates

CountSupport

In dieser Phase wird der Support $sup(c)$ jedes generierten Kandidaten $c \in C_k$ ermittelt. Erfüllt ein Kandidat nicht die geforderte Minimalhäufigkeit, wird er aus der Kandidatenmenge entfernt. Dieser Vorgang lässt sich dadurch begründen, dass der Support der Teilsequenzen S' einer Sequenz S mindestens der Support der Sequenz S aufweisen, vergleiche dazu Definition 5.14.

Zur Durchführung der Experimente wurde der von Bockermann ([Boc07]) implementierte GSP-RapidMiner-Operator verwendet.

6. Ergebnisse

In diesem Kapitel werden sowohl die Versuchsumgebung, als auch die Ergebnisse, die mit den in Kapitel 5 beschriebenen Analyseverfahren erzielt wurden, vorgestellt.

6.1. Versuchsumgebung

Alle im Rahmen der Diplomarbeit durchgeführten Prozesse der Datenspeicherung und -aufbereitung wurden auf dem Microsoft SQL-Server 2005 [SQL] ausgeführt. Die Experimente wurden mit RapidMiner [MWK+06], einem Open-Source System für Wissensentdeckung in Datenbanken und Data Mining, durchgeführt. RapidMiner stellt eine Vielzahl von Operatoren für alle Phasen der Wissensentdeckung in Datenbanken zur Verfügung, vom Einlesen der Daten, Datenverarbeitung, Lernverfahren bis hin zur Validierung. Abbildung 6.1 zeigt ein Beispielexperiment in RapidMiner.

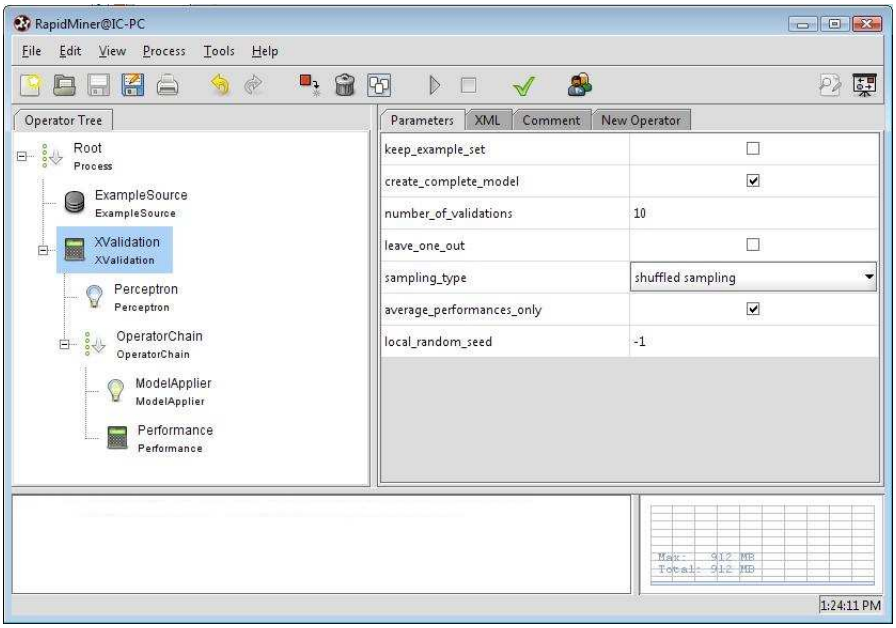


Abbildung 6.1.: Experiment in Rapidminer

6.2. Auswertung der Ergebnisse

Im Folgenden werden nun die Ergebnisse der einzelnen Lernverfahren vorgestellt.

6.2.1. Klassifikationsverfahren

Perceptron

Die Ergebnisse des Perceptrons, welcher in Abschnitt 5.2.1 beschrieben wurde, hängen stark von der gewählten Darstellung der Krankengeschichte und der betrachteten Versichertengruppe ab. Tabelle 6.1 fasst alle Ergebnisse zusammen. *Anmerkung:* Die Ergebnisse mit der höchsten Accuracy werden aufgrund der Übersichtlichkeit **fett** dargestellt.

Trainingsmenge	Accuracy	Precision	Recall	F-Measure
F17-Diag. Vorerkrankungen	58,95%	55,36%	93,51%	34,77%
F17-Diag. Häufigkeit	71,32%	67,48%	82,98%	37,21%
F17-Diag. Häufigkeit + Zeit	74,12%	70,49%	83,16%	38,15%
F17-Diag. Krankheitsdauer	67,81%	67,62%	68,07%	33,92%
F17-Diag. Krankheitsdauer + Zeit	69,39%	66,52%	78,6%	36,02%
F1-Diag. Vorerkrankungen	59,9%	56,01%	93,5%	35,03%
F1-Diag. Häufigkeit	63,9%	59,17%	90,70%	35,81%
F1-Diag. Häufigkeit + Zeit	71,55%	66,08%	89%	37,92%
F1-Diag. Krankheitsdauer	63,45%	60,45%	78,5%	34,15%
F1-Diag. Krankheitsdauer + Zeit	68,5%	66,05%	76,5%	35,46%
F3-Diag. Vorerkrankungen	52,62%	51,26%	67,82%	29,19%
F3-Diag. Häufigkeit	53,47%	52,49%	66,17%	29,27%
F3-Diag. Häufigkeit + Zeit	52,51%	51,31%	98,13%	33,69%
F3-Diag. Krankheitsdauer	58,14%	55,35%	84,39%	33,43%
F3-Diag. Krankheitsdauer + Zeit	65,29%	62,49%	76,45%	34,38%
F4-Diag. Vorerkrankungen	50,45%	50,23%	99,78%	33,41%
F4-Diag. Häufigkeit	50,47%	50,24%	99,66%	33,40%
F4-Diag. Häufigkeit + Zeit	51,54%	50,79%	98,45%	33,50%
F4-Diag. Krankheitsdauer	54,67%	52,71%	91,38%	33,42%
F4-Diag. Krankheitsdauer + Zeit	62,99%	60,07%	77,56%	33,85%
F-Diag. Vorerkrankungen	50,14%	50,07%	99,88%	33,35%
F-Diag. Häufigkeit	50,10%	50,5%	99,72%	33,52%
F-Diag. Häufigkeit + Zeit	50,67%	50,34%	99,15%	33,39%
F-Diag. Krankheitsdauer	54,41%	52,45%	94,63%	33,75%
F-Diag. Krankheitsdauer + Zeit	64,84%	61,36%	80,24%	34,77%

Tabelle 6.1.: Lernergebnisse des Perceptron-Algorithmus

Gute Ergebnisse können bei Suchtkranken erzielt werden (Accuracy zwischen 59,9% und 71,55%). Betrachtet man eine spezielle Untergruppe der Suchtkranken, nämlich die Rau-

cher, so bewegt sich die Accuracy zwischen $58,95\%$ und $74,12\%$, je nach Darstellung der Krankengeschichte. Bei Versicherten mit neurotischen, Belastungs- und somatoformen Störungen, welche die größte Untergruppe der psychisch Kranken bilden, werden die schlechtesten Ergebnisse erzielt (maximale Accuracy beträgt $62,99\%$).

Wie bereits genannt hängt das Ergebnis stark von der Darstellung der Krankenvorgeschichte ab. Bei den Suchtkranken lassen sich im Vergleich zu psychisch unauffälligen Personen in der Häufigkeit des Auftretens von Krankheiten Unterschiede feststellen. Innerhalb der übrigen Versichertengruppen scheint die Dauer einer Erkrankungen eine Rolle zu spielen. Auffällig dabei ist, dass unter Berücksichtigung des Diagnosezeitpunktes bessere Ergebnisse als ohne Angabe des Zeitpunktes erzielt werden.

Versicherte mit psychischen und Verhaltensstörungen durch Tabak Nachstehend werden die fünf charakteristischsten Krankheiten jeder einzelnen Darstellung in tabellarischer Form präsentiert. Dabei ist die Aussagekraft der einzelnen Ergebnisse aus Tabelle 6.1 zu berücksichtigen.

Die Tabelle 6.2 zeigt die identifizierten typischen "Rauchermerkmale".

Merkmal
Alter $>$ 20 Jahre und $<$ 70 Jahre
Schuppenflechte (L40)
Vorzeitige Wehen (O60)
Akute Bronchitis (J20)
Zahnverfall durch systemische Ursachen (K08)

Tabelle 6.2.: Ergebnisse der F17-Vorerkrankung-Darstellung

Die Tabellen 6.3 und 6.4 listen die Krankheiten auf, an denen Raucher häufiger leiden als Versicherte ohne psychische Störung.

Merkmal
Schuppenflechte (L40)
Harnstau (N13)
Fehlbildungen des peripheren Gefäßsystems (Q27)
Vorzeitige Wehen (O60)
Wundrose (A46)

Tabelle 6.3.: Ergebnisse der F17-Häufigkeit-Darstellung

Merkmal	Zeitpunkt [in Jahren]
erbliche Nierenkrankheiten (N07)	$t-3$
Verdauungssystemkrankheiten (K92)	$t-3$
Muskelverletzung im Unterschenkel (S86)	$t-3$
Schuppenflechte (L40)	$t-2$
Gutartige Neubildungen im Verdauungssystems(D13)	$t-3$

Tabelle 6.4.: Ergebnisse der F17-Häufigkeit+Zeit-Darstellung

Anmerkung zur Darstellung der Ergebnisse mit Zeitangaben: Die Zeitangaben sind in Jahren dargestellt und t bezeichnet den Zeitpunkt der ersten Diagnose einer psychischen

Störung. Beispiel: $t-4$ bedeutet, dass die Versicherten vier Jahre vor dem ersten Auftreten einer psychischen Störung an Krankheit xy litten.

Nachfolgend werden in Tabelle 6.5 und 6.6 die Krankheiten dargestellt, an den Raucher länger behandelt wurden als psychisch unauffällige Versicherte.

Merkmal
Schuppenflechte (L40)
Biomechanische Funktionsstörungen (M99)
Zerrung von Gelenken des Schultergürtels (S43)
Fraktur im Bereich der Schulter (S42)
Bösartige Neubildung der Brustdrüse (C50)

Tabelle 6.5.: Ergebnisse der F17-Krankheitsdauer-Darstellung

Merkmal	Zeitpunkt [in Jahren]
Verletzungen der Schulter (M75)	$t-4$
sonst. Bursopathien (M71)	$t-4$
Verbrennung des Handgelenkes (T23)	$t-4$
chronische obstruktive Lungenkrankheiten (J44)	$t-4$
Verrenkung mit Beteiligung mehrerer Körperregionen (T03)	$t-4$

Tabelle 6.6.: Ergebnisse der F17-Krankheitsdauer+Zeit-Darstellung

Auffällig bei allen dargestellten Ergebnissen ist, dass Schuppenflechte bei vier von fünf Darstellungen eine Rolle spielt. So wird Schuppenflechte als eine typische "Raucherkrankheit" identifiziert. Dies wird auch durch drei andere Darstellungen bestätigt. Raucher leiden länger und häufiger an Schuppenflechte als Versicherte ohne eine psychische Störung. Dieses Ergebnis wird auch durch eine Studie aus den USA (vgl. [SCC07]) untermauert. Als typische Raucherkrankheiten wurden ebenfalls Komplikationen in der Schwangerschaft und bei der Geburt, Herzerkrankungen, Lungenkrankheiten, Magengeschwüre und sogar Brustkrebs vom Perceptron erkannt. Diese werden auch von der WHO [WHO02] als typische Raucherkrankheiten genannt.

Versicherte mit psychischen und Verhaltensstörungen durch psychotrope Substanzen
Tabelle 6.7 listet die für Suchtkranke charakteristischen Diagnosen auf.

Merkmal
Nasenschleimhautentzündung (J31)
Schuppenflechte (L40)
sonst. Krankheiten der Nase und -nebenhöhle (J34)
Nasenpolyp (J33)
Nasennebenhöhlenentzündung(J32)

Tabelle 6.7.: Ergebnisse der F1-Vorerkrankung-Darstellung

Auffällig bei diesem Ergebnis ist, dass vier von fünf Diagnosen Krankheiten der Nasenschleimhaut und Nasennebenhöhlen betreffen. Dieser Zusammenhang konnte aber noch

nicht durch andere Studien verifiziert werden. Eine Wechselwirkung zwischen Schuppenflechte und einer Suchtkrankheit lässt sich dadurch erklären, dass Raucher die größte Untergruppe der Suchtkranken bilden. So wurde die typische Raucherkrankheit auch als typische Suchtkrankheit erkannt.

In Tabellen 6.8 und 6.9 werden die Diagnosen aufgezeigt, an denen Suchtkranke häufiger leiden als Versicherte ohne psychische Auffälligkeiten.

Merkmal
angeborene Fehlbildungen des peripheren Gefäßsystems (Q27)
Vorhofflattern und Vorhofflimmern (I48)
chron. Niereninsuffizienz (N18)
Atherosklerose (I70)
Diabetes mellitus (E10)

Tabelle 6.8.: Ergebnisse der F1-Häufigkeit-Darstellung

Merkmal	Zeitpunkt [in Jahren]
Komplikationen eines Traumas (T89)	$t-1$
Halsentzündung (J02)	$t-3$
Migräne (G43)	$t-4$
Krankheiten der Nägel (L60)	$t-1$
Spontangeburt eines Einlings (O80)	$t-1$

Tabelle 6.9.: Ergebnisse der F1-Häufigkeit+Zeit-Darstellung

Auffällig bei der Darstellung, in der nur die Häufigkeit einer Diagnose betrachtet wurde, ist, dass Herz- und Arterienkrankheiten dominieren. Der Zusammenhang zwischen missbräuchlichem Konsum von Alkohol und Tabak mit Herzkrankheiten wird auch in diversen Studien nachgewiesen. Starker Alkoholkonsum über Jahre führt zur Abnahme der Kontraktionskraft des Herzens, einer dosisabhängigen Blutdruckerhöhung, Vorhof- und Kammerarrhythmien und zu einer Vergrösserung des Herzens (vgl. [SM01]). Rauchen hat zudem eine gefäßschädigende Wirkung. Folgen des Rauchens sind der Anstieg der Pulsfrequenz, des koronaren Widerstands, der myokardialen Kontraktilität und des myokardialen Sauerstoffbedarfs ([HM05]).

In der Häufigkeit+Zeit-Darstellung betrifft jede identifizierte Krankheit einen anderen Bereich. Ein Zusammenhang zwischen zwei der identifizierten Diagnosen und einer Suchterkrankung wurde auch in anderen Studien nachgewiesen.

In mehreren Studien ([dK96] und [MY96]) wurde belegt, dass ein sehr enger Zusammenhang zwischen einer Alkohol- oder Drogenabhängigkeit und einem erlittenen Trauma besteht. Der stärkste Leidensdruck eines Menschen mit traumatischen Erfahrungen besteht darin, dass er immer wieder mit seinen schlimmen Erinnerungen konfrontiert wird. Deshalb wird oft die Beruhigung durch Alkohol oder Drogen mit dämpfender Wirkung als Ausweg gewählt.

Zunächst erscheint ein Zusammenhang zwischen einer Suchtkrankheit und einer Geburt befremdlich. Doch jede zehnte Mutter leidet nach der Geburt an einer postnatalen Depression (vgl. [Rho04]). Eine Studie des Centre for Addiction and Mental Health und der University of North Dakota ([GMDR07]) hat herausgefunden, dass Frauen, die unter starken Depressionen leiden, häufig trinken, um einen Ausweg aus ihren Problemen zu finden. Werden diese beiden Erkenntnisse kombiniert, so lässt sich der Zusammenhang

zwischen einer Suchtkrankheit und einer Geburt erklären.

Die Tabellen 6.10 und 6.11 listen die Krankheiten auf, aufgrund derer Suchtkranke länger ärztliche Leistungen in Anspruch genommen haben als psychisch unauffällige Versicherte.

Merkmal
angeborene Fehlbildungen der Extremitäten (Q74)
angeborene Muskel-Skelett-Deformitäten (Q68)
Biomech. Funktionsstörungen (M99)
Kontaktdermatitis (L25)
Virushepatitis (B17)

Tabelle 6.10.: Ergebnisse der F1-Krankheitsdauer-Darstellung

Merkmal	Zeitpunkt [in Jahren]
Zerrung von Gelenken des Schultergürtels (S43)	$t-4$
Fraktur im Bereich der Schulter (S42)	$t-4$
Schulterläsionen (M75)	$t-4$
Sodbrennen (R12)	$t-4$
Bursopathien (M71)	$t-4$

Tabelle 6.11.: Ergebnisse der F1-Krankheitsdauer+Zeit-Darstellung

Versicherte mit affektiven Störungen Unter affektiven Störungen versteht man akute, chronische oder episodische Stimmungsstörungen, wozu u.a. Depressionen und übermäßig gesteigerte Euphorie zählen. Folgende typische Merkmale wurden aus den einzelnen Darstellungen der Krankengeschichte für Versicherte mit affektiven Störungen, also Versicherte mit einer F3-Diagnose, identifiziert:

Merkmal
entzündungsbedingte, knotenartige Gewebeneubildung der Haut (L92)
spezielle Untersuchungen bei Personen ohne Beschwerden (Z01)
Präpartale Blutungen (O46)
Abzess in der Rektalregion (K61)
Bandscheibenschäden (M50)

Tabelle 6.12.: Ergebnisse der F3-Vorerkrankung-Darstellung

Die Wechselwirkung zwischen einer präpartalen Blutung und affektiven Störungen wird als einzige der hier aufgeführten Diagnosen durch andere Studien belegt. Präpartale Blutungen sind Komplikationen während der Schwangerschaft. Dass Frauen während einer Schwangerschaft oder nach der Geburt häufig unter Stimmungsproblemen und Depressionen leiden ([Rho04]), wurde bereits näher erläutert.

Die Diagnose "Abszess in der Rektalregion", die als charakteristische Affektstimmungskrankheit erkannt wurde, wird ebenfalls in den beiden folgenden Darstellungen (siehe Tabelle 6.13 und 6.14) als typische Krankheit von Versicherten mit affektiven Störungen ermittelt. Eine Wechselwirkung zwischen Abszessen in der Rektalregion und Depressionen kann jedoch durch keine weiteren Studien nachgewiesen werden.

Merkmal
Abszess in der Rektalregion (K61)
Bakterien als Krankheitsursache (B96)
spezielle Untersuchungen bei Personen ohne Beschwerden (Z01)
Krankheiten der Wirbelsäule (M53)
Viruswarzen(B07)

Tabelle 6.13.: Ergebnisse der F3-Häufigkeit-Darstellung

Merkmal	Zeitpunkt [in Jahren]
Bakterien als Krankheitsursache (B96)	$t-2$
Kollaps, Ohnmacht (R55)	$t-1$
Abszess in der Rektalregion (K61)	$t-3$
Myeloische Leukämie (C92)	$t-3$
Sonnenbrand (L55)	$t-2$

Tabelle 6.14.: Ergebnisse der F3-Häufigkeit+Zeit-Darstellung

Tabelle 6.14 zeigt, dass ein Zusammenhang zwischen Leukämie und Depressionen gefunden wurde. Die Diagnose von Leukämie als eine Ursache für Depressionen betrachtet werden kann, ist instinktiv nachvollziehbar. Die signifikante Zunahme der Depressivität nach solch einer Diagnose wird ebenfalls [Ins03] wissenschaftlich bestätigt.

Die beiden folgenden Tabellen listen die Krankheiten auf, bei denen Versicherte mit affektiven Störungen länger ärztliche Behandlung in Anspruch nehmen als Versicherte ohne psychische Störungen.

Merkmal
Zystische Nierenkrankheit (Q61)
Krankheiten der Brustdrüse (O92)
Kardiomyopathie (I42)
Osteoporose (M80)
kardiale Arrhythmien (I49)

Tabelle 6.15.: Ergebnisse der F3-Krankheitsdauer-Darstellung

Merkmal	Zeitpunkt [in Jahren]
Knorpelkrankheiten (M94)	$t-2$
Gelenkschädigungen (M24)	$t-2$
Rhizarthrose (M18)	$t-2$
Multiple Sklerose (G35)	$t-2$
kardiale Arrhythmien (I49)	$t-1$

Tabelle 6.16.: Ergebnisse der F3-Krankheitsdauer+Zeit-Darstellung

In beiden Darstellungen wird ein Zusammenhang zwischen Herzkrankheiten und Depressionen ersichtlich. Medizinische Studien bestätigen diese Wechselwirkung ([Gla07]). Die Untersuchungen zeigen, dass 17 bis 27 Prozent der Krankenhauspatienten mit Herzinfarkt oder ähnlichen Herzkrankheiten zugleich schwere Depressionen haben.

Zwei weitere Diagnosen, die vom Perceptron als charakteristisch für Versicherte mit somatoformen Störungen identifiziert wurden, werden durch Studien nachgewiesen.

Zum einen finden Forscher des Karolinska Institutes in Stockholm eine hohe Verbreitung depressiver Symptome unter Multiple Sklerose-Patienten (vgl. [GEF+07]) und zum anderen wird durch eine Studie des National Institute of Mental Health der USA ([EMT+07]) eine Wechselwirkung zwischen Osteoporose und Depressionen belegt .

Versicherte mit neurotischen, Belastungs- und somatoformen Störungen Als somatoforme Störungen werden körperliche Beschwerden bezeichnet, die sich nach gründlicher Untersuchung nicht auf eine organische Erkrankung zurückführen lassen. Charakteristisch für diese Patienten ist die wiederholte Darbietung körperlicher Beschwerden verbunden mit Forderungen nach medizinischen Untersuchungen. Obwohl die Ursachen der Beschwerden nicht im körperlichen Symptomen begründet liegen und die Ärzte die immer wieder beteuern, bestehen die Patienten auf weiterer Untersuchungen. Die körperlichen Symptome, von denen diese Patienten berichten, können sich auf jedes Körperteil oder jedes Körpersystem beziehen. Es ist selbst für Mediziner schwer eine somatoforme Störung als solche zu erkennen (vgl. [SE07]), was sich auch hier in den vorgestellten Ergebnissen widerspiegelt. In der Gruppe der Versicherten mit somatoformen Störungen erzielt das Perceptron im Vergleich zu anderen Versichertengruppen die schlechtesten Ergebnisse (vgl. Tabelle 6.1).

Nachfolgend werden die vom Perceptron erkannten charakteristischen Krankheiten für Versicherte mit neurotischen, Belastungs- und somatoformen Störungen (Versicherte mit einer F4-Diagnose) aufgelistet.

Merkmal
Entzündung der Gallenblase (K81)
Kontaktdermatitis (L25)
Venenkrankheiten (I87)
abnorme Befunde der Blutchemie (R79)
Sonnenbrand (L55)

Tabelle 6.17.: Ergebnisse der F4-Vorerkrankung-Darstellung

In Tabelle 6.18 und 6.19 werden die Diagnosen aufgezeigt, aufgrund derer Patienten mit somatoformen Störungen häufiger einen Arzt konsultierten als andere Patienten.

Merkmal
Verdauungssystemkrankheiten durch med. Maßnahmen (K91)
angeborene Fehlbildungen des Harnsystems (Q64)
bösart. Neubildung des Knochens (C40)
bösart. Neubildung des Eierstocks (C56)
Störungen des Wasser- und Elektrolythaushaltes (E87)

Tabelle 6.18.: Ergebnisse der F4-Häufigkeit-Darstellung

Merkmal	Zeitpunkt [in Jahren]
Krankheiten des Verdauunssystem(K91)	$t-1$
angeborene Fehlbildungen des Harnsystems (Q64)	$t-1$
Mehrlingsschwangerschaft (O30)	$t-3$
Eitrige Gelenkentzündung (M00)	$t-2$
angeborene Fehlbildungen des Gesichts (Q18)	$t-3$

Tabelle 6.19.: Ergebnisse der F4-Häufigkeit+Zeit-Darstellung

Die Tabellen 6.20 und 6.21 listen die Krankheiten auf, bei denen sich die Behandlungsdauer von Versicherten mit somatoformen Störungen von der Behandlungsdauer der Versicherte ohne psychische Störungen unterscheidet.

Merkmal
Spirochäteninfektionen (A69)
Rheumatisches Fieber (I00)
reaktive Arthritiden (M02)
Postinfektiöse und reaktive Arthritiden (M03)
entzündliche Spondylopathien (M46)

Tabelle 6.20.: Ergebnisse der F4-Krankheitsdauer-Darstellung

Merkmal	Zeitpunkt [in Jahren]
Rheumatisches Fieber (I00)	$t-3$
Spirochäteninfektionen (A69)	$t-3$
Juvenile Arthritis (M08)	$t-3$
Postinfektiöse Arthritiden (M03)	$t-3$
Reaktive Arthriden (M46)	$t-3$

Tabelle 6.21.: Ergebnisse der F4-Krankheitsdauer+Zeit-Darstellung

Versicherte mit psychischen und Verhaltensstörungen Diese Versichtengruppe umfasst alle Ausprägungen von psychischen Störungen (insbesondere auch die zuvor dargestellten Versichertengruppen). Nicht verwunderlich dabei ist, dass gewisse Krankheiten, die bereits Unterkategorien zugeordnet wurden, ebenso in dieser Oberkategorie erkannt wurden.

Nachfolgend werden die typischen Merkmale aus den einzelnen Darstellungen der Krankengeschichte von Patienten mit psychischen Störungen (Versicherte mit einer F-Diagnose) gezeigt.

Merkmal
sonst. Venenkrankheiten (I87)
Verstauchung und Zerrung des Ellenbogens (S53)
schwangerschaftsinduzierte Flüssigkeitsansammlungen (O12)
Überwachung einer normalen Schwangerschaft (Z34)
Spontangeburt eines Einlings (O81)

Tabelle 6.22.: Ergebnisse der F-Vorerkrankung-Darstellung

Auffällig bei den Ergebnissen der F-Vorerkrankung-Darstellung ist, dass drei der fünf aufgelisteten Diagnosen Schwangerschaften und Geburten betreffen. Der Zusammenhang zwischen Geburten und Depressionen wurde bereits mehrmals erläutert. Depressionen bilden die zweitgrößte Unterkategorie der Patienten mit psychischen Störungen. Dies ist eine mögliche Erklärung des Ergebnisses.

Kennzeichnend für die folgenden beiden Darstellung (siehe Tabelle 6.23 und 6.23), welche die Häufigkeit einer Diagnose betrachten, sind die onkologischen Befunde (böseartige Neubildungen und Leukämie). Psychische Störungen sind ein häufiges Problem bei onkologischen Patienten, da ca. 5% bis 46% aller Krebspatienten im Verlauf der Erkrankung

unter Depressionen, Anpassungsstörungen, Angststörungen und Verwirrtheitszuständen leiden (vgl. [Sch07] und [Sti04]).

Merkmal
Verdauungssystem- krankheiten durch med. Maßnahmen (K91)
bösart. Neubildung des Rektums (C20)
Leukämie (C91)
bösart. Neubildung der Atmungs- und Verdauungsorgane (C78)
angeb. Fehlbildungen des Harnsystems (Q64)

Tabelle 6.23.: Ergebnisse der F-Häufigkeit-Darstellung

Merkmal	Zeitpunkt [in Jahren]
Verdauungssystemkrankheiten durch med. Maßnahmen (K91)	$t-1$
bösart. Neubildung des Magens (C16)	$t-2$
Leukämie (C91)	$t-1$
bösart. Neubildung an sonst. Lokalisationen (C79)	$t-1$
angeb. Fehlbildungen des Harnsystems (Q64)	$t-1$

Tabelle 6.24.: Ergebnisse der F-Häufigkeit+Zeit-Darstellung

Wird die Krankheitsdauer von Patienten mit psychischen Störungen betrachtet, sind die Krankheitsbilder sehr weit gestreut, vergleiche dazu Tabelle **??** und 6.26.

Merkmal
sonst. Knochen-krankheiten (M89)
rheumatisches Fieber (I00)
Spirochäteninfektion (A69)
Laktoseintoleranz (E73)
Postinfektiöse und reaktive Arthriden (M03)

Tabelle 6.25.: Ergebnisse der F-Krankheitsdauer-Darstellung

Merkmal	Zeitpunkt [in Jahren]
Hakenwurm-Krankheit (B76)	$t-3$
chronische Bronchitis (J41)	$t-3$
angeb. Fehlbildungen der oberen Extremitäten (Q74)	$t-3$
Herzninfarkt (I21)	$t-2$
bösart. Neubildung des Gebärmutterhalses	$t-2$

Tabelle 6.26.: Ergebnisse der F-Krankheitsdauer+Zeit-Darstellung

Naive Bayes

Naive Bayes kann kein Konzept aus den vorliegenden Datensätzen lernen und ist somit für diese Art der Problemstellung nicht geeignet. Dieses Verfahren lieferte keinerlei verwendbare Ergebnisse (Accuracy = 50%), unabhängig von der Art der Darstellung und der Versichertengruppen.

6.2.2. Subgruppenentdeckung

Zur Erkennung lokaler Modelle wird das in Kapitel 5.3 beschriebene Verfahren verwendet. Die Anzahl der lokalen Modelle wurde auf maximal zehn begrenzt.

Perceptron

Auch bei der Suche lokaler Modelle mit Hilfe des Perceptrons, spielt die Darstellung der Krankengeschichte eine entscheidende Rolle. Tabelle 6.27 gibt eine Übersicht über die erzielten Resultate.

Trainingsmenge	Accuracy	Precision	Recall	F-Measure
F17-Diag. Vorerkrankungen	60,09%	60,12%	90,18	36,07%
F17-Diag. Häufigkeit	71,4%	67,42%	83,16%	37,23
F17-Diag. Häufigkeit + Zeit	72,98%	69,7%	81,58%	37,58%
F17-Diag. Krankheitsdauer	71,05%	69,5%	75,26%	36,13%
F17-Diag. Krankheitsdauer + Zeit	70,26%	67,42%	79,12%	36,4%
F1-Diag. Vorerkrankungen	62,55%	60,15%	92,4%	36,43%
F1-Diag. Häufigkeit	65,05%	60,56%	86,5%	35,62%
F1-Diag. Häufigkeit + Zeit	70,75%	65,38%	88,6%	37,62%
F1-Diag. Krankheitsdauer	66,1%	62,22%	81,7%	35,32%
F1-Diag. Krankheitsdauer + Zeit	69,35%	66,73%	77,4%	37,92%
F3-Diag. Vorerkrankungen	56,95%	55,86%	72,06%	31,47%
F3-Diag. Häufigkeit	53,06%	52,48%	64,71%	28,98%
F3-Diag. Häufigkeit + Zeit	54,49%	52,44%	96,81%	34,01%
F3-Diag. Krankheitsdauer	58,87%	56,02%	82,35%	33,34%
F3-Diag. Krankheitsdauer + Zeit	64,9%	62,1%	76,36%	76,36%
F4-Diag. Vorerkrankungen	51,6%	50,83%	98,21%	33,49%
F4-Diag. Häufigkeit	50,57%	50,29%	99,44%	33,39%
F4-Diag. Häufigkeit + Zeit	52,9%	51,55%	96,76%	33,63%
F4-Diag. Krankheitsdauer	54,53%	52,63%	91,09%	33,56%
F4-Diag. Krankheitsdauer + Zeit	63,01%	60,28%	76,33%	33,68%
F-Diag. Vorerkrankungen	51,21%	50,62%	99,08%	33,5%
F-Diag. Häufigkeit	50,28%	50,14%	99,63%	33,35%
F-Diag. Häufigkeit + Zeit	51,73%	50,9%	98,45%	33,55%
F-Diag. Krankheitsdauer	56,14%	53,48%	94,6%	34,17%
F-Diag. Krankheitsdauer + Zeit	64,48%	61,02%	80,22%	33,66%

Tabelle 6.27.: Lernergebnisse der Supgruppenentdeckung mit Perceptron

Vergleicht man diese Ergebnisse mit denen des Perceptron in der Tabelle 6.1, so fällt auf, dass die besten Resultate des Perceptron zur Suche globaler Modelle nicht übertroffen werden. Bei der Analyse der restlichen Darstellungen kann bei der Suche nach lokalen Modellen mit Hilfe des Perceptrons nur eine marginale Verbesserung erreicht werden. Aufgrund des Umfangs wird auf eine detaillierte Präsentation der gefundenen Modelle

verzichtet.

Entscheidungsbaum

Wie aus der Tabelle 6.28 zu entnehmen ist, werden auch hier die besten Ergebnisse bei Rauchern und Suchtkranken erreicht. Im Bereich der Suchtkranken (Versicherte mit einer F1-Diagnose) werden sogar bessere Ergebnisse als beim Perceptron-Algorithmus erzielt. Im Gegensatz zum Perceptron-Algorithmus haben die einzelnen Darstellungen der Krankengeschichten einer Versichertengruppe nicht so einen starken Einfluss auf die Ergebnisse.

Trainingsmenge	Accuracy	Precision	Recall	F-Measure
F17-Diag. Vorerkrankungen	71,93%	68,98%	81,05%	37,26%
F17-Diag. Häufigkeit	71,58%	68,57%	80,00%	36,92%
F17-Diag. Häufigkeit + Zeit	68,77%	67,54%	74,91%	35,52%
F17-Diag. Krankheitsdauer	71,75%	68,43%	81,23%	37,14%
F17-Diag. Krankheitsdauer + Zeit	68,6%	72,08%	66,14%	34,49%
F1-Diag. Vorerkrankungen	71,37%	78,11%	59,6%	33,81%
F1-Diag. Häufigkeit	72,45%	78,87%	61,5%	34,56%
F1-Diag. Häufigkeit + Zeit	72,5%	78,22%	62,4%	34,71%
F1-Diag. Krankheitsdauer	70,5%	77,09%	58,7%	33,32%
F1-Diag. Krankheitsdauer + Zeit	72,75%	78,03%	63,4%	34,71%
F3-Diag. Vorerkrankungen	69,54%	70,58%	67,85%	34,59%
F3-Diag. Häufigkeit	69,56%	71,58%	65,18%	34,16%
F3-Diag. Häufigkeit + Zeit	67,54%	67,28%	68,53%	33,95%
F3-Diag. Krankheitsdauer	68,39%	69,93%	64,84%	33,64%
F3-Diag. Krankheitsdauer + Zeit	66,9%	66,34%	69,47%	33,95%
F4-Diag. Vorerkrankungen	67,9%	68,63%	66,25%	33,71%
F4-Diag. Häufigkeit	68,46%	69,16%	66,93%	34,01%
F4-Diag. Häufigkeit + Zeit	67,01%	70,91%	58,61%	32,09%
F4-Diag. Krankheitsdauer	67,13%	69,11%	62,82%	32,91%
F4-Diag. Krankheitsdauer + Zeit	66,76%	69,66%	59,84%	32,09%
F-Diag. Vorerkrankungen	68,8%	67,1%	74,29%	35,26%
F-Diag. Häufigkeit	68,82%	67,16%	74,16%	35,24%
F-Diag. Häufigkeit + Zeit	68,96%	70,91%	64,4%	33,75%
F-Diag. Krankheitsdauer	68,92%	69,15%	69,12%	34,57%
F-Diag. Krankheitsdauer + Zeit	68,37%	70,57%	63,07%	33,3%

Tabelle 6.28.: Lernergebnisse der Subgruppenentdeckung mit DecisionTree der Tiefe 5

Zunächst wurde die Tiefe des Entscheidungsbaums auf drei begrenzt. Diese Einschränkung lieferte jedoch Regeln, welche sich nur auf Alter, Postleitzahl, Berufsstatus oder Geschlecht bezogen und keine Aussagen über Diagnosen enthielten. Aus diesem Grund wurde die Tiefe des Baumes auf fünf erhöht. Der Nachteil dabei ist, dass die erzeugten Regeln relativ kompliziert sind. Zudem beschreiben die generierten Regeln, im Gegensatz

zu den bisher betrachteten Verfahren, die Diagnosen, an welchen die einzelnen Versichertengruppen nicht erkranken.

Aufgrund des Umfangs werden an dieser Stelle nur die Ergebnisse der einzelnen Versichertengruppen vorgestellt, welche die höchste Accuracy erreichen.

Versicherte mit psychischen und Verhaltensstörungen durch Tabak Die besten Ergebnisse wurden bei der Analyse der Vorerkrankungsübersicht erzielt.

Nachfolgend werden die gefunden Regeln der einzelnen Modelle dargestellt:

Modell 1

1.1	(Alter = 60 bis 69) $\wedge \neg$(M54)	$\rightarrow$ F17 (N=2, P=35)
1.2	(Alter = 40 bis 49) $\wedge \neg$(I10) $\wedge \neg$(D25)	$\rightarrow$ F17 (N=67, P=162)
1.3	(Alter = 50 bis 59) $\wedge \neg$(T14) $\wedge \neg$(M77)	$\rightarrow$ F17 (N=25, P=136)

Modell 2

2.1	(Alter > 9) $\wedge \neg$(J06) $\wedge \neg$(Z38)	$\rightarrow$ F17 (N=379, P=535)
2.2	(Alter > 9) $\wedge$ (J06) $\wedge$ (M62)	$\rightarrow$ F17 (N=0, P=3)

Modell 3

3.1	(Alter > 9) $\wedge$ (weiblich) $\wedge \neg$(J03) $\wedge \neg$(J02)	$\rightarrow$ F17 (N=256, P=334)
3.1	(Alter = 70 bis 79) $\wedge$ (männlich)	$\rightarrow$ F17 (N=0, P=5)
3.3	(Alter = 60 bis 69) $\wedge$ (männlich) $\wedge \neg$(E11)	$\rightarrow$ F17 (N=25, P=136)

Modell 4

4.1	(Alter > 9) $\wedge \neg$(J03) $\wedge$ (R69) $\wedge \neg$(S93)	$\rightarrow$ F17 (N=446, P=566)

Modell 5

5.1	(Alter > 9) $\wedge \neg$(M17) $\wedge \neg$(J11)	$\rightarrow$ F17 (N=489, P=560)

Modell 6

6.1	$\neg$(FAMI)[1] $\wedge \neg$(J30) $\wedge \neg$(I99) $\wedge \neg$(R69)	$\rightarrow$ F17 (N=470, P=497)
6.2	(Alter > 9) $\wedge$ (FAMI) $\wedge \neg$(J30) $\wedge \neg$(O09)	$\rightarrow$ F17 (N=489, P=560)
6.3	(J30) $\wedge$ (J20)	$\rightarrow$ F17 (N=0, P=4)

Modell 7

7.1	$\neg$(FAMI) $\wedge \neg$(J30) $\wedge \neg$(S92) $\wedge \neg$(C53)	$\rightarrow$ F17 (N=483, P=497)
7.2	(FAMI) $\wedge \neg$(J30) $\wedge \neg$(O09) $\wedge$ (R55)	$\rightarrow$ F17 (N=0, P=2)
7.3	(J30)	$\rightarrow$ F17 (N=2, P=5)

Modell 8

8.1	$\neg$(FAMI) $\wedge \neg$(J30) $\wedge \neg$(S92) $\wedge \neg$(C53)	$\rightarrow$ F17 (N=483, P=497)
8.2	(Alter = 50 bis 59) $\wedge$ (FAMI) $\wedge \neg$(J30) $\wedge \neg$(R55)	$\rightarrow$ F17 (N=0, P=21)
8.3	(FAMI) $\wedge \neg$(J30) $\wedge$ (R55)	$\rightarrow$ F17 (N=0, P=2)
8.4	(J30)	$\rightarrow$ F17 (N=2, P=5)

Modell 9

9.1	$\neg$(FAMI) $\wedge \neg$(J30) $\wedge \neg$(S92) $\wedge \neg$(C53)	$\rightarrow$ F17 (N=483, P=497)
9.2	(Alter > 9) $\wedge$ (FAMI) $\wedge \neg$(J30) $\wedge \neg$(O09)	$\rightarrow$ F17 (N=21, P=68)
9.3	(J30)	$\rightarrow$ F17 (N=2, P=5)

[1] FAMI ist eine Abkürzung für Familienmitversicherte.

Modell 10

10.1	$\neg$(FAMI) $\wedge$ $\neg$(J30) $\wedge$ $\neg$(M70) $\wedge$ $\neg$(S50)	$\rightarrow$ F17 (N=480, P=497)
10.2	(Alter = 50 bis 59) $\wedge$ (FAMI) $\wedge$ $\neg$(J30) $\wedge$ $\neg$(R55)	$\rightarrow$ F17 (N=0, P=21)
10.3	(FAMI) $\wedge$ $\neg$(J30) $\wedge$ (R55)	$\rightarrow$ F17 (N=0, P=2)
10.4	(J30)	$\rightarrow$ F17 (N=2, P=5)

Exemplarisch wird zunächst anhand der Regel 2.1 die Notation erläutert. Regel 2.1 besagt, dass Versicherte, die 10 Jahre oder älter sind und keine akute Infektion der oberen Atemwege erleiden und nicht gebären, mit einer erhöhten Wahrscheinlichkeit rauchen. Diese Art von Regeln sind leider nicht intuitiv nachvollziehbar und nur schwer durch Studien zu belegen.

Was anhand der Ergebnisse jedoch sehr deutlich wird, ist die Tatsache, dass 22 der 27 Regeln mindestens eine Aussage über soziodemografische Merkmale (Alter, Beruf, Geschlecht) enthalten. Insbesondere scheint der Versichertenstatus, aus welchem sich ein mögliches Beschäftigungsverhältnis ableiten lässt, in Kombination mit mindestens einer nicht aufgetretenen Krankheit ein Indiz für psychische Störungen durch Tabakkonsum zu sein. Häufig nicht auftretende Krankheiten sind z.B. allergische Rhinopathie[1](J30) und Synkope und Kollaps[2] (R55).

Versicherte mit psychischen und Verhaltensstörungen durch psychotrope Substanzen
Bei Versicherten mit einer F1-Diagnose wurden die besten Ergebnisse bei der Analyse der Krankengeschichte in Form der Krankheitsdauer+Zeitangabe-Übersicht erzielt. Nachfolgend werden die einzelnen Modell beschrieben.

Modell 1

1.1	(Alter = 40 bis 49) $\wedge$ $\neg$(FAMI)	$\rightarrow$ F1 (N=147, P=240)
1.2	(Alter $>$ 9) $\wedge$ $\neg$(Alter = 50 bis 59) $\wedge$ $\neg$(Z37, t-2)	$\rightarrow$ F1 (N=39, P=151)
1.3	(Alter = 50 bis 59) $\wedge$ $\neg$(FAMI) $\wedge$ $\neg$(T14, t-3) $\wedge$ $\neg$(M77, t-3)	$\rightarrow$ F1 (N=39 , P=165)
1.4	(Alter = 50 bis 59) $\wedge$ (FAMI) $\wedge$ $\neg$(T14, t-3)	$\rightarrow$ F1 (N=39 , P=165)

Modell 2

2.1	(Alter $>$ 9) $\wedge$ $\neg$(Alter = 60 bis 79) $\wedge$ $\neg$(J06, t-3)	$\rightarrow$ F1 (N=147, P=240)
2.2	(Alter = 70 bis 79)	$\rightarrow$ F1 (N=1 , P=22)
2.3	(Alter = 60 bis 69) $\wedge$ $\neg$(M54, t-3) $\wedge$ $\neg$(M54, t-2)	$\rightarrow$ F1 (N=5, P=56)

Modell 3

3.1	(Alter = 70 bis 79)	$\rightarrow$ F1 (N=1, P=22)
3.2	(Alter = 60 bis 69)	$\rightarrow$ F1 (N=9, P=49)

Modell 4

4.1	$\neg$(J06, t-2) $\wedge$ $\neg$(K08, t-2) $\wedge$ $\neg$(T14, t-3) $\wedge$ $\neg$(J02, t-3)	$\rightarrow$ F1 (N=885, P=982)
4.2	(J06, t-2) $\wedge$ (T14, t-3 $>$ 75,5 Tage)	$\rightarrow$ F1 (N=0, P=2)
4.3	(männlich) $\wedge$ (J06, t-2) $\wedge$ (A09, t-1) $\wedge$ $\neg$(J30, t-3) $\wedge$ $\neg$(J02, t-3)	$\rightarrow$ F1 (N=0 , P=2)
4.4	(J06, t-2) $\wedge$ (J30, t-2)	$\rightarrow$ F1 (N=0, P=2)

Modell 5

5.1	(Alter = 50 bis 59) $\wedge$ $\neg$(FAMI) $\wedge$ $\neg$(N99, t-3) $\wedge$ $\neg$(K29, t-2)	$\rightarrow$ F1 (N=44, P=165)
5.2	(Alter = 50 bis 59) $\wedge$ (FAMI)	$\rightarrow$ F1 (N=0, P=29)

[1] allergischer Schnupfen
[2] Blackout und Ohnmacht

Modell 6

6.1	$\neg$(R69, t-4) $\wedge$ $\neg$(M54, t-2) $\wedge$ $\neg$(Z38, t-3) $\wedge$ $\neg$(Z38, t-2)	$\rightarrow$ F1 (N=917, P=993)
6.2	$\neg$(R69, t-4) $\wedge$ (M54, t-2) $\wedge$ (Z38, t-3 $>$ 2 Tage)	$\rightarrow$ F1 (N=0, P=29)

Modell 7

7.1	$\neg$(M54, t-3) $\wedge$ $\neg$(D62, t-2) $\wedge$ $\neg$(Z37, t-1) $\wedge$ $\neg$(O62, t-2)	$\rightarrow$ F1 (N=920, P=971)
7.2	(M54, t-3) $\wedge$ (M51, t-3)	$\rightarrow$ F1 (N=1, P=5)

Modell 8

8.1	$\neg$(O62, t-2) $\wedge$ $\neg$(D62, t-2) $\wedge$ $\neg$(K08, t-2) $\wedge$ $\neg$(S60, t-3)	$\rightarrow$ F1 (N=969, P=995)
8.2	$\neg$(O62, t-2) $\wedge$ (D62, t-2)	$\rightarrow$ F1 (N=0, P=3)
8.3	(O62, t-2) $\wedge$ (O09, t-2 $\geq$ 3 Tage)	$\rightarrow$ F1 (N=1, P=2)

Modell 9

9.1	(Alter $>$ 9) $\wedge$ $\neg$(O62, t-2) $\wedge$ $\neg$(D62, t-2) $\wedge$ $\neg$(B34, t-2)	$\rightarrow$ F1 (N=903, P=992)
9.2	$\neg$(O62, t-2) $\wedge$ (D62, t-2)	$\rightarrow$ F1 (N=0, P=3)
9.3	(O62, t-2) $\wedge$ (O09, t-2 $\geq$ 3 Tage)	$\rightarrow$ F1 (N=1, P=2)

Modell 10

10.1	$\neg$(O62, t-2) $\wedge$ $\neg$(D62, t-2) $\wedge$ $\neg$(T14, t-3) $\wedge$ $\neg$(J02, t-3)	$\rightarrow$ F1 (N=952, P=993)
10.2	$\neg$(O62, t-2) $\wedge$ $\neg$(D62, t-2) $\wedge$ (T14, t-3 $\geq$ 75, Tage)	$\rightarrow$ F1 (N=0, P=3)
10.3	(O62, t-2) $\wedge$ (O09, t-2 $<=$ 3 Tage)	$\rightarrow$ F1 (N=1, P=2)

Die an dieser Stelle verwendete Notation soll anhand der Regel 1.4 verdeutlicht werden. Die Regel 1.4 besagt, dass Familienmitversicherte im Alter zwischen 50-59 Jahren, die vor drei Jahren nicht an Sehnentzündungen und Verletzungen an einer Körperregion litten, mit einer erhöhten Wahrscheinlichkeit Suchtkrank sind.

Betrachtet man die Regeln der ersten drei Modelle, wird deutlich, dass das Alter der Versicherten ein Indiz für Suchterkrankungen ist. Insbesondere Versicherte im Alter zwischen 40 bis 69 scheinen anfälliger für Suchterkrankungen zu sein.

Ebenso auffällig bei Regeln der Modelle 8 bis 10 ist, dass abnorme Wehentätigkeit (O62) und akute Blutungsanämie[3](D62) im gleichen Zeitraum bei suchtkranken Versicherten nicht auftreten.

Versicherte mit affektiven Störungen Bei Versicherten mit affektiven Störungen wurden die besten Ergebnisse bei der Analyse der Vorerkrankungsübersicht erreicht. Hierbei konnten allerdings nur sechs Subgruppen identifiziert werden.

Modell 1

1.1	(weiblich) $\wedge$ (Alter $>$ 9) $\wedge$ $\neg$(Alter = 20 bis 29)	$\rightarrow$ F3 (N=623, P=1429)
1.2	(Alter $>$ 9) $\wedge$ (Pflichtmitglied)	$\rightarrow$ F3 (N=0, P=2)
1.3	(männlich) $\wedge$ (Alter $>$ 9) $\wedge$ (T14) $\wedge$ $\neg$(B01)	$\rightarrow$ F3 (N=0 , P=2)

Modell 2

2.1	(Alter $>$ 9) $\wedge$ $\neg$(Alter = 30 bis 39) $\wedge$ $\neg$(J06) $\wedge$ $\neg$(Z38)	$\rightarrow$ F3 (N=1049, P=1493)
2.2	(Alter $>$ 9) $\wedge$ (Pflichtmitglied)	$\rightarrow$ F3 (N=0, P=2)
2.3	(FAMI) $\wedge$ $\neg$(J06) $\wedge$ $\neg$(Z30) $\wedge$ $\neg$(Z38)	$\rightarrow$ F3 (N=0 , P=3)

[3]Blutarmut

Modell 3

3.1	(Alter $>$ 9) $\wedge$ $\neg$(R69) $\wedge$ $\neg$(T14) $\wedge$ $\neg$(Z38)	$\rightarrow$ F3 (N=1785, P=2226)
3.2	(Alter $>$ 9) $\wedge$ $\neg$(R69) $\wedge$ (T14) $\wedge$ (M70)	$\rightarrow$ F3 (N=0, P=5)

Modell 4

4.1	(Alter $>$ 29) $\wedge$ $\neg$(S93)	$\rightarrow$ F3 (N=1239, P=1811)
4.2	(Alter = 20 bis 29) $\wedge$ (K52) $\wedge$ (R07)	$\rightarrow$ F3 (N=0, P=2)

Modell 5

5.1	$\neg$(K08) $\wedge$ $\neg$(J20) $\wedge$ $\neg$(K52 $\wedge$ $\neg$(M54))	$\rightarrow$ F3 (N=1601, P=1924)
5.2	$\neg$(K08) $\wedge$ (J20) $\wedge$ $\neg$(J01) $\wedge$ (N92)	$\rightarrow$ F3 (N=0, P=5)
5.3	$\neg$(K08) $\wedge$ (J20) $\wedge$ (J01) $\wedge$ (T81)	$\rightarrow$ F3 (N=0, P=2)
5.4	(K08) $\wedge$ $\neg$(T14) $\wedge$ (T81)	$\rightarrow$ F3 (N=0, P=2)
5.5	(K08) $\wedge$ (T14) $\wedge$ (J06)	$\rightarrow$ F3 (N=2, P=5)

Modell 6

6.1	(Alter $>$ 19) $\wedge$ $\neg$(T15)$\neg$(S52)	$\rightarrow$ F3 (N=2076, P=2293)
6.2	$\neg$(Alter = 10 bis 19) $\wedge$ (S52) $\wedge$ (I84) $\wedge$ $\neg$(T15)	$\rightarrow$ F3 (N=0, P=2)

Um nochmals zu verdeutlichen, wie schwer es ist Korrektheit der generierten Regeln nach zu vollziehen, werden exemplarisch diejenigen mit der größten Abdeckung jedes einzelnen Modells beschrieben.

Regel 1.1 besagt, dass weibliche Versicherte, im Alter zwischen 10 und 19 oder die älter als 30 Jahre sind, an affektiven Störungen leiden. Aus Regel 2.1 kann abgeleitet werden, dass Versicherte im Alter zwischen 10 und 29 oder die älter als 40 Jahre sind und weder entbunden haben noch an einer akuten Infektion der oberen Atemwege erkrankten, leiden mit einer erhöhten Wahrscheinlichkeit an affektiven Störungen. Aus Regel 3.1 ist zu entnehmen, dass Versicherte, die älter als 9 Jahre sind und weder entbunden haben, noch eine Verletzung erlitten und nicht an einer akuten Infektion der oberen Atemwege erkrankten, mit einer erhöhten Wahrscheinlichkeit an affektiven Störungen leiden. Regel 4.1 besagt, dass Versicherte, die älter als 30 Jahre sind und sich nicht den Fuß verstaucht haben, an Depressionen leiden. Aus Regel 5.1 ist zu entnehmen, dass Versicherte, welche weder an Krankheiten der Zähne, noch an Rückschmerzen, noch an nichtinfektiöser Gastroenteritis erkranken, mit einer erhöhten Wahrscheinlichkeit an affektiven Störungen leiden. Regel 6.1 drückt aus, dass Versicherte die älter als 20 Jahre sind und weder einen Fremdkörper im Augen hatten, noch eine Fraktur des Unterarms erlitten, mit einer erhöhten Wahrscheinlichkeit an affektiven Störungen leiden.

Versicherte mit neurotischen, Belastungs- und somatoformen Störungen Bei Versicherten mit einer F4-Diagnose wurden die besten Ergebnisse bei der Analyse der Krankengeschichte in Form der Krankheitsdauer+Zeitangabe-Übersicht erzielt. Nachfolgend werden die Regeln der einzelnen Modell dargestellt.

Modell 1

1.1	(männlich) $\wedge$ (Alter $>$ 9) $\wedge$ $\neg$(Alter = 20 bis 29) $\wedge$ $\neg$(Z38)	$\rightarrow$ F4 (N=987, P=2333)
1.2	(männlich) $\wedge$ (Alter = 20 bis 29) $\wedge$ (Z38) $\wedge$ (O12)	$\rightarrow$ F4 (N=0, P=2)
1.3	(weiblich) $\wedge$ $\neg$(Alter = 40 bis 49) $\wedge$ (J06) $\wedge$ (J32)	$\rightarrow$ F4 (N=5, P=9)

Modell 2

2.1 (Alter > 9) $\wedge \neg$(Alter = 30 bis 39) $\wedge \neg$(K08) $\wedge \neg$(J06) $\rightarrow$ F4 (N=1848, P=2431)
2.2 (Alter = 30 bis 39) $\wedge \neg$(J06) $\wedge \neg$(R69) $\rightarrow$ F4 (N=1042, P=1151)
2.3 (Alter = 30 bis 39) $\wedge$ (J06) $\wedge$ (O60) $\rightarrow$ F4 (N=2, P=7)

Modell 3

3.1 (Alter > 9) $\wedge \neg$(Alter = 30 bis 39) $\wedge \neg$(Z38) $\wedge \neg$(T14) $\rightarrow$ F4 (N=2121, P=2674)
3.2 (Alter = 40 bis 49) $\wedge$ (Z38) $\wedge$ (Z35) $\rightarrow$ F4 (N=0, P=5)
3.3 (Alter = 30 bis 39) $\wedge$ (J00) $\wedge$ (S93) $\wedge \neg$(G47) $\rightarrow$ F4 (N=1, P=3)
3.4 (Alter = 30 bis 39) $\wedge$ (E66) $\wedge$ (G47) $\rightarrow$ F4 (N=2, P=22)

Modell 4

4.1 $\neg$(M54) $\wedge \neg$(J03) $\wedge \neg$(S93) $\wedge \neg$(J20) $\rightarrow$ F4 (N=2919, P=3434)
4.2 $\neg$(M54) $\wedge \neg$(J03) $\wedge$ (S93) $\wedge$ (M22) $\rightarrow$ F4 (N=0, P=4)
4.3 $\neg$(M54) $\wedge$ (J03) $\wedge \neg$(J06) $\wedge$ (I63) $\rightarrow$ F4 (N=2, P=8)
4.4 (M54) $\wedge \neg$(J06) $\wedge$ (I63) $\rightarrow$ F4 (N=0, P=3)
4.5 (M54) $\wedge$ (J06) $\wedge \neg$(A09) $\wedge$ (J35) $\rightarrow$ F4 (N=0, P=3)
4.6 $\neg$(Alter = 30 bis 39) $\wedge$ (M54) $\wedge$ (J06) $\wedge$ (A09) $\rightarrow$ F4 (N=7, P=15)

Modell 5

5.1 $\neg$(Alter = 10 bis 19) $\wedge \neg$(J06) $\wedge \neg$(R69) $\wedge \neg$(S60) $\rightarrow$ F4 (N=3129, P=3537)
5.2 $\neg$(Alter = 10 bis 19) $\wedge \neg$(J06) $\wedge$ (O70) $\rightarrow$ F4 (N=0, P=2)
5.3 (Alter = 10 bis 19) $\wedge$ (männlich) $\wedge \neg$(J06) $\wedge$ (K35) $\rightarrow$ F4 (N=47, P=70)
5.4 (Alter = 10 bis 19) $\wedge$ (weiblich) $\wedge \neg$(Z03) $\wedge \neg$(K35) $\rightarrow$ F4 (N=40, P=53)

Modell 6

6.1 $\neg$(Alter = 10 bis 29) $\wedge \neg$(R69) $\wedge \neg$(J35) $\rightarrow$ F4 (N=2431, P=3051)
6.2 $\neg$(Alter = 10 bis 29) $\wedge$ (R69) $\wedge$ (B99) $\rightarrow$ F4 (N=1, P=4)
6.3 (Alter = 10 bis 19) $\wedge$ (männlich) $\wedge \neg$(J06) $\wedge$ (K35) $\rightarrow$ F4 (N=47, P=70)
6.4 (Alter = 10 bis 19) $\wedge$ (weiblich) $\wedge \neg$(Z03) $\wedge \neg$(K35) $\rightarrow$ F4 (N=40, P=53)

Modell 7

7.1 (Alter ≥ 19) $\wedge$ (männlich) $\wedge \neg$(J00) $\rightarrow$ F4 (N=2058, P=3111)
7.2 $\neg$(Alter = 10 bis 19) $\wedge$ (männlich) $\wedge$ (J00) $\wedge$ (B34) $\rightarrow$ F4 (N=3, P=7)
7.3 $\neg$(Alter = 10 bis 29) $\wedge$ (weiblich) $\wedge$ (T88) $\rightarrow$ F4 (N=0, P=5)
7.4 (Alter = 10 bis 19) $\wedge$ (weiblich) $\wedge \neg$(K35) $\wedge \neg$(J06) $\rightarrow$ F4 (N=47, P=70)
7.5 (Alter = 10 bis 19) $\wedge$ (weiblich) $\wedge \neg$(Z03) $\wedge \neg$(S52) $\rightarrow$ F4 (N=40, P=53)

Modell 8

8.1 $\neg$(Alter = 10 bis 19) $\wedge \neg$(G47) $\wedge \neg$(S60) $\wedge \neg$(R69) $\rightarrow$ F4 (N=3827, P=3910)
8.2 $\neg$(Alter = 10 bis 19) $\wedge \neg$(FAMI) $\wedge$ (G47) $\rightarrow$ F4 (N=15, P=52)
8.3 (Alter = 10 bis 19) $\wedge$ (männlich) $\wedge \neg$(K35) $\wedge \neg$(J06) $\rightarrow$ F4 (N=47, P=70)
8.4 (Alter = 10 bis 19) $\wedge$ (weiblich) $\wedge \neg$(K35) $\wedge \neg$(Z03) $\rightarrow$ F4 (N=40, P=53)

Modell 9

9.1 $\neg$(Alter = 10 bis 19) $\wedge \neg$(J06) $\wedge \neg$(N83) $\wedge \neg$(T15) $\rightarrow$ F4 (N=3212, P=3550)
9.2 $\neg$(Alter = 10 bis 19) $\wedge \neg$(J06) $\wedge$ (N83) $\wedge$ (O80) $\rightarrow$ F4 (N=0, P=2)
9.3 (Alter = 10 bis 19) $\wedge$ (männlich) $\wedge \neg$(K35) $\wedge \neg$(S50) $\rightarrow$ F4 (N=50, P=70)
9.4 (Alter = 10 bis 19) $\wedge$ (weiblich) $\wedge \neg$(K35) $\wedge \neg$(S52) $\rightarrow$ F4 (N=42, P=53)

Modell 10

10.1	$\neg$(Alter = 10 bis 19) $\wedge$ $\neg$(G47) $\wedge$ $\neg$(N83) $\wedge$ $\neg$(S61)	$\rightarrow$ F4 (N=3914, P=3917)
10.1	$\neg$(Alter = 10 bis 19) $\wedge$ $\neg$(G47) $\wedge$ (N83) $\wedge$ (O00)	$\rightarrow$ F4 (N=0, P=2)
10.3	$\neg$(Alter = 10 bis 19) $\wedge$ (G47)	$\rightarrow$ F4 (N=20, P=52)
10.4	(Alter = 10 bis 19) $\wedge$ (männlich) $\wedge$ $\neg$(K35) $\wedge$ $\neg$(S50)	$\rightarrow$ F4 (N=50, P=70)
10.5	(Alter = 10 bis 19) $\wedge$ (weiblich) $\wedge$ $\neg$(Z03) $\wedge$ $\neg$(S52)	$\rightarrow$ F4 (N=40, P=53)

Obwohl bisher die Korrektheit der generierten Regeln schwer nachvollziehbar war, zeigt Regel 1.1, dass Zusammenhänge richtig erfasst werden. In Regel 1.1 wird der Zusammenhang zwischen Männern und dem Merkmal "nicht schwanger" entdeckt.

Bei Versicherten mit somatoformen Störungen ist das Alter ein wiederkehrendes Kriterium, insbesondere das Alter zwischen 10 und 19 Jahren in Kombination mit dem nicht Auftreten einer akuten Blinddarmentzündung (K35).

Versicherte mit psychischen und Verhaltensstörungen Bei Versicherten mit einer F-Diagnose wurden die besten Ergebnisse bei der Analyse der Krankengeschichte in Form der Krankheitsdauer+Zeitangabe-Übersicht erzielt. Nachfolgend werden die Regeln der einzelnen Modell dargestellt.

Modell 1

1.1	(weiblich) $\wedge$ (Alter $>$ 9) $\wedge$ $\neg$(Alter = 20 bis 29) $\wedge$ (PLZ $\neq 0x$)[1]	$\rightarrow$ F (N=1957, P=4064)
1.2	(PLZ = $0x$) $\wedge$ (N70, t-2)	$\rightarrow$ F (N=0, P=2)
1.3	(PLZ = $0x$) $\wedge$ $\neg$(N70, t-2) $\wedge$ (G43, t-1)	$\rightarrow$ F (N=1, P=2)

Modell 2

2.1	(Alter = 30 bis 39) $\wedge$ (PLZ $\neq 1x$) $\wedge$ (PLZ $\neq 6x$) $\wedge$ (PLZ $\neq 9x$)	$\rightarrow$ F (N=4294, P=5352)
2.2	(Alter = 30 bis 39) $\wedge$ (PLZ = $1x$) $\wedge$ (M70, t-1)	$\rightarrow$ F (N=1, P=2)

Modell 3

3.1	(PLZ $\neq 1x$) $\wedge$ (PLZ $\neq 3x$) $\wedge$ (PLZ $\neq 6x$) $\wedge$ $\neg$(J06, t-3)	$\rightarrow$ F (N=5854, P=7149)

Modell 4

4.1	$\neg$(Alter $>$ 9) $\wedge$ (PLZ $\neq 3x$) $\wedge$ (PLZ $\neq 6x$) $\wedge$ $\neg$(J06, t-3)	$\rightarrow$ F (N=5854, P=7149)

Modell 5

5.1	$\neg$(Alter = 20 bis 29) $\wedge$ $\neg$(R69, t-4) $\wedge$ $\neg$(M54, t-3) $\wedge$ $\neg$(Z38, t-3)	$\rightarrow$ F (N=4381, P=5664)
5.2	(Alter = 20 bis 29) $\wedge$ (J20, t-3) $\wedge$ $\neg$(G43, t-1)	$\rightarrow$ F (N=0, P=2)

Modell 6

6.1	$\neg$(T14, t-3) $\wedge$ $\neg$(J20, t-3) $\wedge$ $\neg$(Z38, t-1) $\wedge$ $\neg$(M54, t-4)	$\rightarrow$ F (N=6596, P=7199)
6.2	$\neg$(T14, t-3) $\wedge$ $\neg$(J20, t-3) $\wedge$ (Z38, t-1) $\wedge$ (D68, t-1)	$\rightarrow$ F (N=0, P=2)
6.3	$\neg$(T14, t-3) $\wedge$ (J20, t-3) $\wedge$ $\neg$(J06, t-3) $\wedge$ (G43, t-1)	$\rightarrow$ F (N=0, P=3)
6.4	(T14, t-3) $\wedge$ $\neg$(J98, t-3) $\wedge$ (J40, t-2)	$\rightarrow$ F (N=0, P=4)
6.5	(T14, t-3) $\wedge$ (J98, t-3)	$\rightarrow$ F (N=0, P=2)

Modell 7

7.1	(PLZ $\neq 3x$) $\wedge$ (PLZ $\neq 4x$) $\wedge$ (PLZ $\neq 6x$) $\wedge$ $\neg$(T14, t-4)	$\rightarrow$ F (N=6864, P=7454)

Modell 8

8.1	(PLZ $\neq 7x$) $\wedge$ $\neg$(Z38, t-2) $\wedge$ $\neg$(R69, t-4) $\wedge$ $\neg$(J03, t-3)	$\rightarrow$ F (N=6903, P=7326)
8.2	(PLZ $\neq 7x$) $\wedge$ (Z38, t-2) $\wedge$ (A09, t-2)	$\rightarrow$ F (N=2, P=3)

[1] PLZ = $0x$ bedeutet, dass die Postleitzahl mit 0 beginnt.

Modell 9

9.1	$(\text{PLZ} \neq 8x) \wedge \neg(\text{J06, t-3}) \wedge \neg(\text{K52, t-3})$	$\rightarrow$ F (N=6537, P=7144)
9.2	$(\text{PLZ} = 8x) \wedge \neg(\text{J03, t-3}) \wedge \neg(\text{K52, t-3}) \wedge (\text{J20, t-1})$	$\rightarrow$ F (N=2, P=3)
9.3	$\neg(\text{J06, t-3}) \wedge (\text{K52, t-3}) \wedge \neg(\text{J20, t-3}) \wedge (\text{N94, t-2})$	$\rightarrow$ F (N=0, P=3)
9.4	$\neg(\text{J06, t-3}) \wedge (\text{K52, t-3}) \wedge (\text{J20, t-3}) \wedge (\text{M54, t-1})$	$\rightarrow$ F (N=1, P=5)
9.5	$(\text{J06, t-3}) \wedge \neg(\text{J20, t-3}) \wedge (\text{K08, t-3}) \wedge (\text{J40, t-3})$	$\rightarrow$ F (N=3, P=8)

Modell 10

10.1	$(\text{PLZ} \neq 8x) \wedge \neg(\text{J06, t-3}) \wedge \neg(\text{K52, t-3})$	$\rightarrow$ F (N=6537, P=7144)
10.2	$(\text{PLZ} = 8x) \wedge \neg(\text{J03, t-3}) \wedge (\text{J20, t-1})$	$\rightarrow$ F (N=0, P=2)
10.3	$(\text{J06, t-3}) \wedge \neg(\text{J20, t-3}) \wedge (\text{K08, t-3}) \wedge (\text{J40, t-3})$	$\rightarrow$ F (N=3, P=8)

Auffällig bei diesen Ergebnissen ist, dass die Region, in denen die Versicherten leben eine Rolle spielt. So besagt Regel 1.1, dass Frauen, die älter als 9 Jahre aber nicht im Alter zwischen 20 und 29 Jahren sind, welche in einer Region wohnen, deren Postleitzahl nicht mit 0 beginnt, mit einer erhöhten Wahrscheinlichkeit an psychischen Störungen leiden.

6.2.3. Entdeckung häufiger Sequenzen in Datenbanken

An dieser Stelle werden die Ergebnisse des GSP Algorithmus vorgestellt, welcher in Kapitel 5.4 vorgestellt wurde.

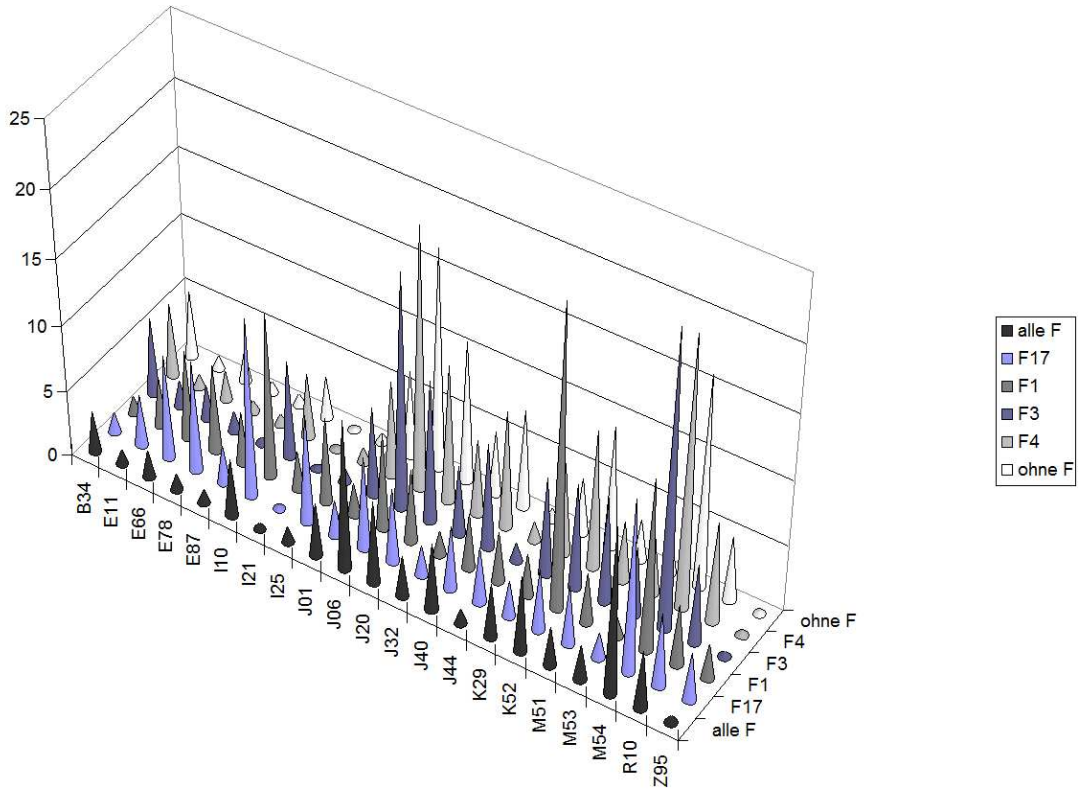


Abbildung 6.2.: Ergebnisse des GSP-Algorithmus

Der GSP-Algorithmus sucht nach häufigen Sequenzen in den Krankengeschichten der Versicherten. Wie bereits in Kapitel 4.3.2 beschrieben, ist eine durchschnittliche Krankengeschichte nur zwischen 4 und 6 Diagnosen lang. Dies ist auch der Grund, weshalb der GSP-Algorithmus nur häufige Sequenzen der Länge 1 findet. Abbildung 6.2 zeigt die Ergebnisse des GSP-Algorithmus. Dargestellt werden nur die Diagnosen, die bei Versicherten mit beliebigen psychischen Störung häufiger auftreten als bei Versicherten ohne eine psychische Störung.

Versicherte mit psychischen und Verhaltensstörungen durch Tabak Tabelle 6.29 listet die Diagnosen auf, bei denen Raucher die größten Unterschiede zu Versicherten ohne psychische Störungen aufweisen. Alle dargestellten Diagnosen werden durch medizinische Studien bestätigt, wobei sechs der sieben Diagnosen sehr eng zusammenhängen. Dies nachfolgend näher erläutert wird.

Support		Diagnose
F17	ohne F	
3,97%	1,01%	Diabetes mellitus [Typ-2-Diabetes] (E11)
7,94%	2,38%	Übergewicht (E66)
8,53%	0,89%	Fettwechselstörung (E78)
13,69%	3,23%	Bluthochdruck (I10)
8,33%	0,73%	ischämische Herzkrankheit (I25)
4,36%	0,91%	chronisch obstruktive Lungenkrankheit (J44)
3,77%	0,21%	kardiale und vaskuläre Implantate (Z95)

Tabelle 6.29.: Häufige Diagnosen der F17-Versicherten

Der deutlich erkennbare Zusammenhang zwischen Übergewicht und Rauchen, sorgt zunächst für Verwirrung. Viele Raucher hoffen mit dem Rauchen ihr Gewicht zu kontrollieren. Jedoch zeigt eine Auswertung mehrerer Studien ([WBG+07]), dass dies ein Irrtum ist, da starke Raucher eher zu Übergewicht neigen. Außerdem verursacht Nikotin eine Insulinresistenz, wodurch das Risiko für das Metabolische Syndrom[4] erhöht wird. Das Metabolische Syndrom steigert wiederum das Risiko einer Herz-Kreislauf-Erkrankung.

Eine weitere Studie [ACM02] bestätigt zudem die Verbindung zwischen Rauchern und der chronisch obstruktiven Lungenkrankheit (COPD)[5]. Dänische Forscher fanden in einer Langzeitstudie heraus, dass die Gefahr an COPD zu erkranken steigt, je länger jemand raucht.

[4] Metabolisches Syndrom - auch tödliches Quartett genannt - umfasst folgende vier Symptome: Übergewicht, Zuckerkrankheit, Fettstoffwechselstörungen und Bluthochdruck

[5] COPD steht für „Chronic Obstructive Pulmonary Disease“. Die chronisch obstruktive Lungenkrankheit umfasst die chronische Entzündung der Bronchien und die daraus resultierende Verengung der Atemwege mit der chronischen Überblähung der Lunge. Wichtigste Krankheitszeichen sind chronischer Husten mit oder ohne Auswurf und eine fortschreitende Luftnot. Die Krankheit ist nicht heilbar.

Versicherte mit psychischen und Verhaltensstörungen durch psychotrope Substanzen Folgende drei Diagnosen treten am häufigsten bei Suchtkranken auf:

Support		Diagnose
F1	ohne F	
4, 04%	0, 87%	Störungen des Elektrolyhaushaltes (E87)
2, 97%	0, 16%	Herzinfarkt (I21)
23%	9, 36%	nichtinfektiöser Brechdurchfall (K52)

Tabelle 6.30.: Häufige Diagnosen der F1-Versicherten

23% aller Suchtkranken klagen im Verlauf ihrer Krankengeschichte über Brechdurchfall. Im Gegensatz dazu leiden nur 9, 36% der Versicherten ohne psychische Störungen an Durchfall und Erbrechen. Brechdurchfall kann durch Bakterien, Viren oder aber durch Vergiftungen verursacht werden ([Roc05]). Bakterien und Viren können aber ausgeschlossen werden, da es sich um einen nicht infektiösen Brechdurchfall handelt. Eine Vergiftung kann u.a. durch Überdosierungen mit legalen und illegalen Drogen, Alkohol oder Medikamenten verursacht werden. Dies könnte die Wechselwirkung zwischen einer Suchterkrankung und dem Brechdurchfall erklären.

Versicherte mit affektiven Störungen Nachfolgend werden die Diagnosen aufgelistet an denen depressive Patienten häufiger als die übrigen hier betrachteten Versichertengruppen erkranken.

Support		Diagnose
F3	ohne F	
5, 97%	5, 15%	Viruskrankheiten (B34)
10, 91%	10, 86%	aktue Bronchitis (J20)
7, 55%	4, 293%	Gastritis und Doudentis (K29)
8, 09%	3, 2%	Bandscheibenschäden (M51)
5, 36%	3, 29%	Krankheiten der Wirbelsäule (M53)
22, 59%	16, 21%	Rückenschmerzen (M54)
6, 18%	4, 99%	Bauch- und Beckenschmerzen (R10)

Tabelle 6.31.: Häufige Diagnosen der F3-Versicherten

Besonders auffällig an diesen Ergebnissen sind die Rückenbeschwerden. Zwar sind Rückschmerzen eine Volkskrankheit, dennoch sind hier signifikante Unterschiede von Patienten mit affektiven Störungen zu Patienten ohne psychischen Störungen zu erkennen. Studien haben gezeigt, dass chronische Rückenbeschwerden und Deperessionen häufig gemeinsam auftreten (vgl. [Dum00]). Daraus geht hervor, dass 80 bis 90 Prozent aller Patienten mit chronischen Rückenschmerzen gleichzeitig leichte depressive Symptome aufweisen. Zudem haben Patienten, bei denen eine Depression diagnostiziert wurde,ein vier Mal höheres Risiko, Rückenschmerzen auszubilden als Patienten ohne Depressionen. Warum Depressionen das Auftreten von Schmerzen begünstigen, ist jedoch noch unklar.

Versicherte mit neurotischen, Belastungs- und somatoformen Störungen Tabelle 6.32 enthält die Diagnosen, an denen Patienten mit somatoformen Störungen häufiger erkranken als die restlichen Versichertengruppen.

Support		Diagnose
F4	ohne F	
7,37%	6,65%	Akute Sinitus (J01)
19,86%	16,81%	Akute Entzündung des Kehlkopfes (J06)
5,79%	4,93%	chronische Sinitus (J32)
9,052%	7,59%	Bronchitis (J40)

Tabelle 6.32.: Häufige Diagnosen der F4-Versicherten

Wie bereits beschrieben ist es selbst für Mediziner sehr schwierig, somatoforme Störungen als solche zu erkennen. Zunächst müssen alle anderen Ursachen ausgeschlossen werden. So müssen sich die Diagnosen der Krankengeschichte dieser Patienten nicht zwangsläufig von psychisch unauffälligen Patienten unterscheiden. Lediglich die Häufigkeit oder Dauer des Auftretens einer Diagnose in der Krankengeschichte dieser Patienten kann eventuell auf eine somatoforme Störung hindeuten. Aus diesem Grund lassen sich bei den Diagnosen der F4-Versicherten nicht so signifikante Unterschiede feststellen wie bei den übrigen Versichertengruppen.

Versicherte mit psychischen und Verhaltensstörungen Obwohl in den Untergruppen, insbesondere bei Rauchern, Süchtigen und depressive Patienten, signifikante Unterschiede zu Versicherten ohne psychische Störungen gefunden wurden, sind in der Gruppe "Versicherte mit psychischen Verhaltensstörungen", die alle Untergruppen beinhaltet, keine großen Unterschiede zu psychisch unauffälligen Versicherten erkennbar. Die Ergebnisse der Untergruppen gehen in der großen Gruppe unter.

7. Zusammenfassung und Ausblick

In diesem Kapitel werden die in dieser Diplomarbeit erzielten Ergebnisse noch einmal zusammengefasst und es wird ein Ausblick auf zukünftige Arbeiten und Entwicklungen gegeben.

7.1. Zusammenfassung

Das Ziel dieser Arbeit ist es, Muster in der Krankengeschichte von psychisch Kranken zu identifizieren, um Versicherten, die erst einen Teil einer typischen Krankengeschichte durchlaufen haben, einen Therapievorschlag und Präventionsmaßnahmen anbieten zu können. Dadurch soll ihnen eine lang andauernde Krankengeschichte erspart und ihre Lebensqualität gesteigert werden. Noch immer sind psychische Krankheiten ein sehr sensibles Thema in unserer Gesellschaft. Ein großer Teil der Bevölkerung hält lieber Distanz zu psychisch Kranken. So haben viele Betroffene Angst vor einer Stigmatisierung. Um an einen potentiell gefährdeten Versicherten herantreten zu können und ihm ein Präventionsprogramm zu unterbreiten, muss beinahe zu 100% sichergestellt sein, dass dieser Versicherte wirklich in absehbarer Zeit an einer psychischen Erkrankung leiden wird.

Diese hohe Wahrscheinlichkeit wird von keinem in dieser Arbeit untersuchten Lernverfahren erreicht.

Dennoch liefert der Perceptron-Algorithmus gute Ergebnisse. Einige davon werden sogar durch medizinische Studien belegt. Bei Suchtkranken (Versicherte mit einer F1 bzw. F17-Diagnose) und der Betrachtung der Häufigkeit des Auftretens einer Erkrankung liefert der Perceptron-Algorithmus die besten Ergebnisse. Wird zu dieser Darstellung noch die Information des Zeitpunktes des Auftretens der Krankheit hinzugenommen, so wird das Ergebnis deutlich verbessert (Steigerung der Accuracy von $63,45\%$ auf $71,55\%$ bei F1-Versicherten bzw. bei F17-Versicherten von $71,32\%$ auf $74,12\%$). Als typische Raucherkrankheiten wurden u.a. Schuppenflechte, die chronisch obstruktive Lungenkrankheit und Herzkrankheiten erkannt. Ein Zusammenhang konnte auch zwischen Trauma bzw. Geburt und der Abhängigkeit von psychotropen Substanzen hergestellt werden. Bei Versicherten mit affektiven (F3), somatoformen (F4) oder allgemein psychischen Störungen (F) liefert das Perceptron die besten Ergebnisse, wenn der Zeitpunkt und die Dauer der einzelnen Krankheiten untersucht wird. Dabei konnte ein Zusammenhang zwischen Schwangerschaft bzw. Geburt und Depressionen nachgewiesen werden. Ebenso konnte eine Wechselwirkung zwischen schweren Krankheiten, wie Leukämie, Herzinfarkt und Multiple Sklerose, und Depressionen nachgewiesen werden.

Naive Bayes hingegen konnte aus den vorliegenden Daten, unabhängig von Versichertengruppen und der Darstellung der Krankengeschichte, kein Konzept lernen, das ein

Objekt einer Klasse zuordnet. Deshalb ist dieses Verfahren für diese Aufgabenstellung ungeeignet.

Im Bereich der Entdeckung von Subgruppen wurde KBS mit zwei unterschiedlichen Basislernern verwendet. Zum einen wurde das Perceptron und zum anderen der DecisionTree genutzt. Das Perceptron konnte an dieser Stelle jedoch die besten Resultate des Perceptron-Algorithmus, welches zur Suche globaler Modelle verwendet wurde, nicht übertreffen. Bei der Analyse der restlichen Darstellungen wurde bei der Suche nach lokalen Modellen mit Hilfe des Perceptrons nur eine marginale Verbesserung der Ergebnisse erzielt. DecisionTree konnte die bereits guten Ergebnisse des Perceptron-Algorithmus im Bereich der Suchtkranken übertreffen. Jedoch sind die vom DecisionTree generierten Regeln relativ komplex und nicht intuitiv nachvollziehbar. Aus diesem Grund lassen sich die Ergebnisse nur schwer durch medizinische Studien belegen.

Der GSP-Algorithmus, der nach häufigen Sequenzen in Datenbanken sucht, konnte aufgrund der relativ kurzen[1] Krankengeschichten der Versicherten nur häufige Sequenzen der Länge eins entdecken. Jedoch werden diese entdeckten Diagnosen auch durch medizinische Studien bestätigt. Der GSP-Algorithmus fand einen Zusammenhang zwischen Rauchern und dem Metabolischen Syndrom, sowie Rauchern und der chronisch obstruktiven Lungenkrankheit heraus. Ebenso wurde entdeckt, dass 23% der Suchtkranken an nicht infektiösem Brechdurchfall leiden. Zum Vergleich dazu leiden nur $9,36\%$ der Versicherten ohne eine psychische Störungen an dieser Krankheit. Ein weiterer Zusammenhang konnte zwischen Depressionen und Rückenbeschwerden hergestellt werden. Zwar sind Rückenschmerzen eine Volkskrankheit[2] aber dennoch ist ein signifikanter Unterschied erkennbar. $22,59\%$ der Patienten, bei denen eine Depression nachgewiesen werden kann, leiden auch unter Rückenschmerzen.

7.2. Ausblick

Im Rahmen dieser Diplomarbeit wurden relativ große psychische Krankheitsgruppen betrachtet, jedoch fasst eine Krankheitsgruppe eine Vielzahl von Störungen unterschiedlichen Schweregrades und mit verschiedenen klinischen Erscheinungsbildern zusammen. Das Beispiel der Suchtkranken (Versicherte mit einer F1-Diagnose) und der Unterkategorie Raucher (Versicherte mit einer F17-Diagnose) legt nahe, dass bessere Ergebnisse erzielt werden können, je präziser die Krankheitskategorie ist. Dies sollte in zusätzlichen Experimenten überprüft werden.

Die betrachtete Krankenkasse hat eine sehr spezielle Versichertenstruktur. Der Großteil der Versicherten ist weiblich und übt einen medizinischen Beruf aus. Personen dieser Gruppen neigen häufiger als andere Gruppen zu psychischen Störungen. Aus diesem Grund sollten zusätzliche Experimente mit Daten anderer Krankenversicherungen durchgeführt werden.

Die Vorläufer späterer psychischer Störungen lassen sich bis ins Kindesalter zurückverfol-

[1] Die durchschnitliche Länge einer Krankengeschichte beträgt fünf Diagnosen.
[2] $16,21\%$ der Versicherten ohne psychische Störungen leiden unter Rückenschmerzen.

gen. Zum jetzigen Zeitpunkt können jedoch nur Daten der letzten fünf Jahre zur Analyse herangezogen werden. Deshalb sollten die hier durchgeführten Experimente zu einem späteren Zeitpunkt wiederholt werden.

Die Datenlage wird sich zudem wahrscheinlich durch die Einführung der elektronischen Gesundheitskarte (eGK) verbessern. Mit der Einführung der elektronischen Gesundheitskarte und dem Aufbau einer Informations- und Kommunikationsinfrastruktur soll eine Verbesserung der Kommunikation aller Beteiligten erreicht werden. Rezepte und Abrechungsdaten sollen zukünftig nur noch elektronisch erstellt und übermittelt werden, so werden die bisherigen fehlerbehafteten Medienbrüche vermieden (vgl. [eGK08]). Auch aus diesem Grund sollten die in dieser Arbeit vorgestellten Experimente zu einem späteren Zeitpunkt wiederholt werden.

Zusammenfassend lässt sich sagen, dass diese Arbeit eine gute Ausgangslage für weitere Experimente bietet. Diese sollten zu einem späteren Zeitpunkt, wenn sich die Datenlage verbessert, durchgeführt werden.

Literaturverzeichnis

[ACM02] Anthouisen, N.R., J.E. Connett und R.P. Murray: *Smoking and Lung Function of Lung Health Study Participants after 11 Years.* American Journal of Respiratory and Critical Care Medicine, 166:675–679, 2002.

[AOK05] *Fehlzeiten-Report 2005: Arbeitsplatzunsicherheit und Gesundheit.* Wissenschaftliches Institut der AOK, 2005.

[AOP08] *Katalog ambulant durchführbarer Operationen und sonstiger stationsersetzender Eingriffe gemäß § 115 b SGB V im Krankenhaus.* 2008. http://www.kbv.de/2613.html.

[AS95] Agrawal, R. und R. Srikant: *Mining sequential patterns.* Research Report RJ 9910, IBM Almaden Research Report, San Jose, California, 1995.

[BFOS84] Breiman, L., J.H. Friedman, R.A. Olshen und C.J. Stone: *CART: Classification and Regression Trees.* Wadsworth, 1984.

[BKK05] *BKK Gesundheitsreport 2005, Krankheitsentwicklungen - Blickpunkt: Psychische Gesundheit.* BKK Bundesverband, 2005. http://www.bkk.de/bkk/show.php3?id=855&nodeid=855.

[Boc07] Bockermann, C. Diplomarbeit, Technische Universität Dortmund, 2007.

[Bol04] Bollmann, D.: *Abrechnung vertragsärztlicher Leistungen, Fremdkassenzahlungsausgleich und Honorarverteilung.* Fortbildungsheft 12 der Kassenärztlichen Bundesvereinigung, 2004. http://www.kbv.de/publikationen/114.html.

[BRDM97] Betz, E., K. Reutter und H. Ritter D. Mecke: *Biologie des Menschen, MörikeBetzMergenthaler.* Quelle& Meyer Verlag, 1997.

[CCK+00] Chapman, P., J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer und R. Wirth: *CRISP-DM 1.0.* Technischer Bericht, http://www.crisp-dm.org/CRISPWP-0800.pdf, 2000. The CRISP-DM Consortium.

[DAK05] *DAK-Gesundheitsreport.* Deutsche Angestelltenkrankenkasse - Versorgungsmanagement, 2005.

[DAL08] *Vertrag über den Datenaustausch auf Datenträgern.* Juli 2008.

[DH73] Duda, R.O. und P.E. Hart: *Pattern Classification and Scene Analysis.* John Wiley & Sons Inc, 1973.

[dK96] Kolk, B.A. van der: *The Complexity of Adaptation to Trauma: Self-Regulation, Stimulus Discrimination, and Characterological Development.*

In: Kolk, B. Van der, A. McFarlane und L. Weisaeth (Herausgeber): *Traumatic Stress: the effects of overwhelming experience on mind, body, and society*, Seiten 182–213. Guilford Press, New York, 1996.

[Dum00] Dumat, W. und Nutzinger, D. O.: *Depressive Entwicklungen bei chronischen Schmerzerkrankungen.* MMW - Fortschritte der Medizin, 3:115–118, 2000.

[eGK08] *Die Gesundheitskarte.* Bundesministerium für Gesundheit, April 2008. http://www.die-gesundheitskarte.de/.

[EMT+07] Eskandari, F., P.E. Martinez, S. Torvik, T.M. Phillips, E.M. Sternberg, S. Mistry, D. Ronsaville, R. Wesley, C. Toomey, N.G. Sebring, J.C. Reynolds, M.R. Blackman, K.A. Calis, P.W. Gold und G. Cizza: *Low Bone Mass in Premenopausal Women With Depression.* Archives of Internal Medicine, 167:2329–2336, 2007.

[FPSS96] Fayyad, U., G. Piatetsky-Shapiro und P. Smyth: *From data mining to knowledge discovery in databases.* Ai Magazine, 17:37–54, 1996. citeseer.ist.psu.edu/fayyad96from.html.

[GEF+07] Gottberg, K., U. Einarsson, S. Fredrikson, L. von Koch und L.W. Holmqvist: *A population-based study of depressive symptoms in multiple sclerosis in Stockholm county: association with functioning and sense of coherence.* Journal of Neurology, Neurosurgery, and Psychiatry, 78:60–65, 2007.

[Gla07] Glassman, AH.: *Depression and cardiovascular comorbidity.* Dialogues in Clinical Neuroscience, 9:9–17, 2007.

[GMDR07] Graham, K., A. Massak, A. Demers und J. Rehm: *Does the Association Between Alcohol Consumption and Depression Depend on How They Are Measured?* Alcoholism: Clinical and Experimental Research, 31:78–88, 2007.

[GPEP07] Giersiepen, K., H. Pohlabeln, G. Egidi und I. Pigeot: *Auszug aus dem Gutachten des Bremer Instituts für Präventionsforschung und Sozialmedizin zur Qualität der Datengrundlagen für morbiditätsbezogene Regelleistungsvolumen in der vertragsärztlichen Versorgung gemäß §§ 85a und 85b SGB V.* Bremer Institut für Präventionsforschung und Sozialmedizin (BIPS), April 2007. http://www.kbv.de/themen/10760.html.

[Gre06] Grethler, A.: *Fachkunde für Kaufleute im Gesundheitswesen.* Thieme, 2006.

[HHB+05] Heidenreich, R., W. Himmel, H. Böckmann, E. Hummers-Pradier, M.M. Kochen, W. Niebling, A. Rogausch, J. Sigle, D. Wetzel und C. Scheidt-Nave: *Elektronische Erfassung von medizinischen Daten in deutschen Hausarztpraxen: Ein Telefon-survey.* Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen, 99:573—-580, 2005.

[HK06] Han, J. und M. Kamber: *Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems).* Morgan Kaufmann, 2006.

[HM05] Heitzer, T. und T. Meinertz: *Rauchen und koronare Herzkrankheit.* Zeitschrift für Kardiologie, 94:iii30–iii42, 2005.

[ICD06] *Internationale statistische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme, 10. Revision - German Modification Version 2007.* Deutsches Institut für Medizinische Dokumentation und Information (DIMDI), Oktober 2006. http://www.dimdi.de/static/de/klassi/diagnosen/icd10/index.htm.

[IELS04] Ihle, W., G. Esser, M. Laucht und M.H. Schmidt: *Depressive Störungen und aggresiv-dissoziale Störungen im Kindes- und Jugendalter.* Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz, 47:728–735, 2004.

[Ins03] Inselmann, U.: *Depressivität bei Patienten mit akuter Leukämie oder hochmalignem Non-Hodgkin-Lymphom.* Doktorarbeit, Institut für Psychotherapie und Medizinische Psychologie, 2003.

[JKW04] Jacobi, F., M. Klose und H.-U. Wittchen: *Psychische Störungen in der Bevölkerung: Inanspruchnahme von Gesundheitsleistungen und Ausfalltage.* Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz, 47:736–744, 2004.

[JW02] Jacobi, F. und H.-U. Wittchen: *Die Versorgungssituation psychischer Störungen in Deutschland.* Psychotherapeutenjournal, 0:6–15, 2002.

[Mit97] Mitchell, T.M.: *Machine Learning.* McGraw-Hill Science/Engineering/Math, 1997.

[MP88] Minsky, M. und S. Papert: *Perceptrons.* Seiten 157–169, 1988.

[MWK+06] Mierswa, I., M. Wurst, R. Klinkenberg, M. Scholz und T. Euler: *YALE: Rapid prototyping for complex data mining tasks.* 2006. citeseer.ist.psu.edu/mierswa06yale.html.

[MY96] McFarlane, A.C. und R. Yehuda: *Resilience, Vulnerability and the Course of Posttraumatic Reactions.* In: Kolk, B.A. Van der, A. McFarlane und L. Weisaeth (Herausgeber): *Traumatic Stress: the effects of overwhelming experience on mind, body, and society*, Seiten 155–181. Guilford Press, New York, 1996.

[Pay02] Payk, T.R.: *Pathopsychologie: Vom Symptom zur Diagnose.* Springer, 2002.

[Pet05] Peterson, H.: *Data Mining - Verfahren, Prozesse, Anwendungsarchitekur.* Oldenbourg, 2005.

[Qua07] Quasdorf, I.: *Aufgaben und Organisation ärztlicher Körperschaften und Verbände.* Fortbildungsheft 1 der Kassenärztlichen Bundesvereinigung, 2007. http://www.kbv.de/publikationen/114.html.

[Rho04] Rhode, A.: *Rund um die Geburt eines Kindes: Depressionen, Ängste und andere psychische Probleme: Ein Ratgeber für Betroffene, Angehörige und ihr soziales Umfeld.* 2004.

[Roc05] *Roche Lexikon Medizin.* Elsevier GmbH, Urban & Fischer Verlag, 2005.

[Ros58] Rosenblatt, F.: *The perceptron: a probabilistic model for information storage and organization in the brain.* Psychological Review, 65:386–408, 1958.

[SA96] Srikant, R. und R. Agrawal: *Mining Sequential Patterns: Generalizations and Performance Improvements.* In: Apers, Peter M. G., Mokrane Bouzeghoub und Georges Gardarin (Herausgeber): *Proc. 5th Int. Conf. Extending Database Technology, EDBT*, Band 1057, Seiten 3–17. Springer-Verlag, 25–29 1996.

[SCC07] Setty, Arathi R., Gary Curhan und Hyon K. Choi: *Smoking and the Risk of Psoriasis in Women: Nurses' Health Study II.* The American Journal of Medicine, 120:953–959, 2007.

[Sch05] Scholz, M.: *Sampling-based sequential subgroup mining.* In: Grossman, Robert, Roberto Bayardo und Kristin P. Bennett (Herausgeber): *KDD*, Seiten 265–274. ACM, 2005.

[Sch07] Schäfer, M.: *Depressionen bei Patienten mit Tumorerkrankungen.* Der Onkologe, 13:632–641, 2007.

[SE07] Sauer, N. und W. Eich: *Somatoforme Störungen und Funktionsstörungen.* Deutsches Ärzteblatt, 104 (1-2):A 45–54, 2007.

[SGB08a] *Fünftes Buch Sozialgesetzbuch - Gesetzliche Krankenversicherung.* 2008. http://bundesrecht.juris.de/sgb_5/index.html.

[SGB08b] *Sechstes Buch Sozialgesetzbuch - Rentenversicherung.* 2008. http://bundesrecht.juris.de/sgb_6/index.html.

[SM01] Schoppet, M. und B. Maisch: *Alkohol und Herz.* Herz 26, 5, 2001.

[SQL] *Microsoft SQL-Server.* Microsoft. http://www.microsoft.com/sql/.

[Stö95] Stöcker, H.: *Taschenbuch mathematischer Formeln und moderner Verfahren.* Verlag Harri Deutsch, 1995.

[Sti04] Stiefel, F.: *Depression und Verwirrtheit bei Krebs.* Im Focus Onkologie, 11:51–54, 2004.

[TK005] *Gesundheitsreport-Auswertungen 2005 zu Trends bei Arbeitsunfähigkeiten und Arzneiverordnungen.* Techniker Krankenkasse, 2005.

[vR79] Rijsbergen, C. J. van: *Information Retrival.* Butterworths, 1979.

[WBG+07] Willi, C., P. Bodenmann, W.A. Ghali, P.D. Faris und J. Cornuz: *Active Smoking and the Risk of Type 2 Diabetes.* JAMA - The Journal of the American Medical Association, 298:2654–2664, 2007.

[WE01] Witten, I.H. und E.Frank: *Data Mining - Praktische Werkzeuge und Technickeen für das maschinelle Lernen.* HANSER, 2001.

[WHO02] WHO: *Machen wir uns nichts vor: Rauchen killt.* Presse-Info EUROPA, 2002.

[WMJ03] Wrobel, S., K. Morik und T. Joachims: *Maschinelles Lernen und Data*

Mining. In: Görz, Günther, C.-R. Rollinger und J. Schneeberger (Herausgeber): *Handbuch der Künstlichen Intelligenz*, Seiten 512–597. Oldenbourg Wissenschaftsverlag, Oldenbourg, 2003.

Erklärung

Hiermit erkläre ich, Alice Czerniejewski, die vorliegende Diplomarbeit mit dem Titel *Analyse von Krankenversichertendaten zur Identifikation psychischer Krankheiten* selbständig verfasst und keine anderen als die hier angegebenen Hilfsmittel verwendet, sowie Zitate kenntlich gemacht zu haben.

Dortmund, 15. Juli 2008